RICE UNIVERSITY

Robust Quantile Regression Using L_2E

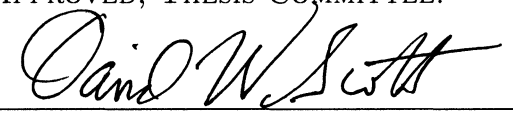
by

Jonathan W. Lane

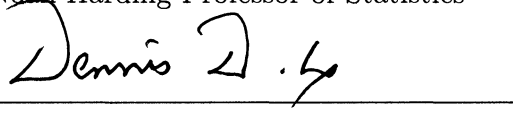
A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

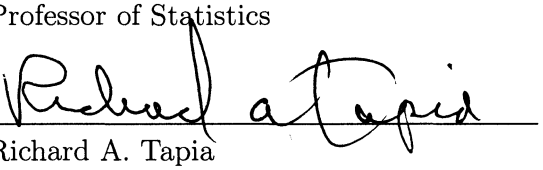
Approved, Thesis Committee:



David W. Scott, Chair
Noah Harding Professor of Statistics



Dennis D. Cox
Professor of Statistics



Richard A. Tapia
University Professor and
Maxfield-Oshman Professor of
Computational and Applied Mathematics

Houston, Texas

November, 2011

ABSTRACT

Robust Quantile Regression Using L_2E

by

Jonathan W. Lane

Quantile regression, a method used to estimate conditional quantiles of a set of data (X, Y), was popularized by Koenker and Bassett (1978). For a particular quantile q, the qth quantile estimate of Y given $X = x$ can be found using an asymmetrically-weighted, absolute-loss criteria. This form of regression is considered to be robust, in that it is less affected by outliers in the data set than least-squares regression. However, like standard L_1 regression, this form of quantile regression can still be affected by multiple outliers. In this thesis, we propose a method for improving robustness in quantile regression through an application of Scott's L_2 Estimation (2001). Theoretic and asymptotic results are presented and used to estimate properties of our method. Along with simple linear regression, semiparametric extensions are examined. To verify our method and its extensions, simulated results are considered. Real data sets are also considered, including estimating the effect of various factors on the conditional quantiles of child birth weight, using semiparametric quantile regression to analyze the relationship between age and personal income, and assessing the value distributions of Major League Baseball players.

Acknowledgments

This project would not have been possible without the help of many individuals. I would first like to thank my committee for their insight and guidance in this paper as well as the knowledge they provided in their classes. In particular, I'd like to thank my advisor, Dr. Scott, who has been the best resource I could ever ask for in this process. This thesis would not have been possible without him and his support. I'd also like to the Department of Statistics at Rice University, and especially Dr. Ensor, for their support and the opportunities they gave me throughout my time with them. I am grateful for all the friends I made through the department, including Eric Chi, David Kahle, Thomas McDonald, Margaret Poon, Beth Bower, and, of course, Team Shark (Stephanie Hicks, Ricardo Affinito, Joe Egbulefu, and Tahira Mammen).

I would also like to thank the Eugene McDermott Scholars Program at the University of Texas at Dallas. In particular, I'd like to acknowledge Dr. Charles Leonard and Sherry Marek who acted as mentors and friends throughout my time with them. I could not be where I am today without all they have done for me. I would also like to thank my friends from the program, especially John McLean, Abraham Rivera, Caitlin Sutton, and Zac Cox.

I would like to give special thanks to my high school calculus teacher, Don Smith from the Albuquerque Academy. He was the man who instilled in me a love of mathematics and the desire to pursue that love that I still carry with me.

I would like to thank my family for all their love and support, namely my parents, Walter and Kathy, and my siblings, Zachary and Jessica. Finally, I owe a great deal to my wife, Alyssa, for her love and encouragement, which have been invaluable to me as I have worked on this paper.

Contents

Illustrations

Tables

Dedication

To Mom, Dad, and Alyssa.

Chapter 1

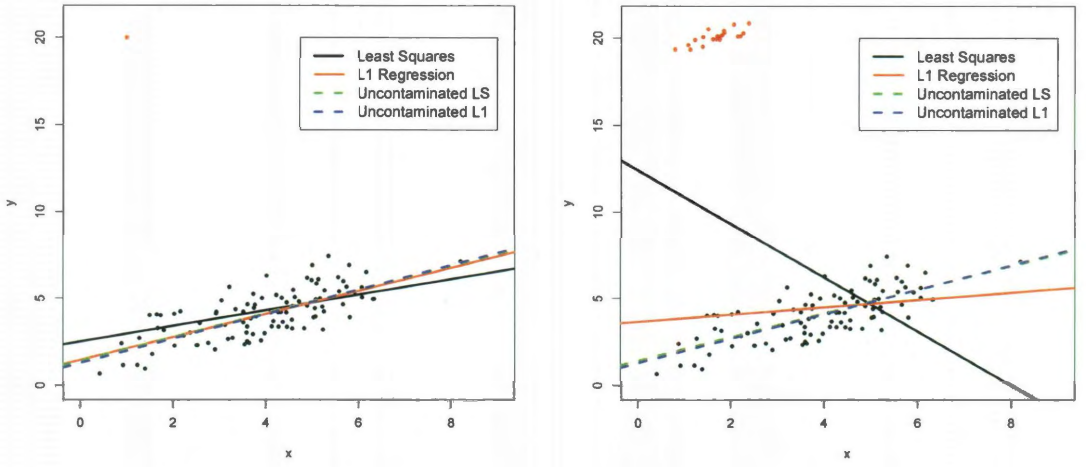
Introduction and Background

Quantile regression, the estimation of the quantiles of a conditional distribution, is a relatively new form of regression that has seen use in several applications in which estimating the distribution of a population is of interest. The prominent form is a generalization of L_1 regression, that is, median regression and thus shares some of the same attributes. One of the popular attributes of median regression shared by this form of quantile regression is robustness, in that the estimate is not greatly affected by extreme outliers unlike ordinary least squares regression.

The differences in the effect that an outlier can have on these two forms of regression are apparent in Figure 1.1(a), in which 99 points of bivariate normal data act as the uncontaminated data, in that there are no added outliers, and one extreme outlier is placed at $(1, 20)$. As we can see, while there is little change in the L_1 regression lines between using the full data set and the uncontaminated data, the least squares regression line is noticeably different.

However, situations can arise where L_1 regression, and thus the prominent form of quantile regression, is affected by multiple outliers. In Figure 1.1(b), the same 99 points of uncontaminated data as before are plotted, but now with a cluster of 31 extreme outliers. Now, not only is the least squares regression line affected by the outliers, the L_1 regression line is greatly affected as well. In situations as these, neither regression method provides a good summary of the uncontaminated data.

In this paper, we propose a robust quantile regression method using L_2 estimation



(a) Without 1 Outlier

(b) With 31 Outliers

Figure 1.1 : Comparison of least squares regression and L_1 regression on data with extreme outliers. Regression lines for the uncontaminated data are also shown.

to be used in these sorts of situations. This chapter provides a background to both quantile regression and to L_2 estimation, both of which serve as the foundation to our method.

1.1 Quantile Regression

Just as least squares regression can be thought of as providing an estimate of the conditional mean of Y given $X = x$, quantile regression can be thought of as providing an estimate of a quantile of Y given $X = x$. That is, for a desired quantile level $\tau \in (0, 1)$, it provides an estimate of the τth quantile, $\hat{\theta}_\tau$, of the conditional distribution of Y given $X = x$.

1.1.1 Koenker and Bassett's Approach

Although there are several methods, perhaps the most well known method of estimating conditional quantiles is the method of quantile regression introduced by Koenker and Basset (1978). The idea stems from the L_1 loss criterion, that is, absolute loss. It is well known that for a sample $(x_1, x_2, \ldots, x_n)$ from a population X, the solution to the minimization problem

$$\arg\min_{\theta} \sum_{i=1}^{n} |x_i - \theta|$$

is $\hat{\theta} = x_{.50}$, that is, the optimal value of θ is the median of the sample. It is also known that we can use this criterion function to perform median regression. So, for example, if we wanted to find the coefficients β in a simple linear model that estimate the conditional median of Y on X, we would minimize the criterion

$$\arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i\beta|.$$

Koenker and Bassett, hereafter referred to as KB, showed that by taking an asymmetrically-weighted absolute loss criterion, rather than the previous symmetric absolute loss criterion, other sample quantiles optimize the resulting minimization problem. Due to its appearance, an example of which is illustrated in Figure 1.3(b), Koenker calls this criterion function the "check" function. To see this, we define the check function for $\tau \in (0, 1)$ by

$$\rho_\tau(x) = \begin{cases} -(1-\tau)x & \text{if } x < 0 \\ \tau x & \text{if } x \geq 0. \end{cases} \tag{1.1}$$

Then, for a sample $(x_1, x_2, \ldots, x_n)$ from X, we solve the minimization

$$\arg\min_{\theta} \sum_{i=1}^{n} \rho_\tau(x_i - \theta), \tag{1.2}$$

which is equivalent to:

$$\arg\min_{\theta} \left[\sum_{x_i<\theta} -(1-\tau) * (x_i - \theta) + \sum_{x_i \geq \theta} \tau \cdot (x_i - \theta) \right]. \tag{1.3}$$

In order to find the optimal value of θ, we look at the first derivative with respect to θ and find its root. Define n_θ to be the number of observations in the sample that have a value less than θ:

$$\begin{aligned} \sum_{x_i<\theta} (1-\tau) * (-1) + \sum_{x_i \geq \theta} \tau &= 0 \\ -n_\theta * (1-\tau) + (n - n_\theta) * \tau &= 0 \\ -n_\theta + n * \tau &= 0. \end{aligned}$$

Thus, the optimal value of θ is the value such that $n_\theta = \tau * n$, that is, the value of θ that lies above $\tau * n$ values of the sample. In other words, $\hat{\theta} = x_\tau$, the τth quantile.

Just as the L_1 loss criterion extends to median regression, the weighted absolute loss criterion extends to quantile regression. For a regression function $g_\theta(x)$ with parameter vector θ, if we minimize the residual error using a particular ρ_τ criterion function, that is,

$$\arg\min_{\theta} \sum_{i=1}^{n} \rho_\tau(y_i - g_\theta(x_i)),$$

we obtain an estimate of the τth conditional quantile. As an example, for simple

linear regression in two dimensions with a sample $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ from (X, Y), if we use a linear regression function $g(x) = \alpha + x\beta$, our minimization becomes

$$\arg\min_{\alpha,\beta} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - x_i\beta).$$

In Figure 1.2(a), examples of this simple linear quantile regression can be seen on 900 points multivariate normal data. In particular, the estimated .01, .05, .1, .25, .5, .75, .9, .95, and .99 quantile regression lines are shown.

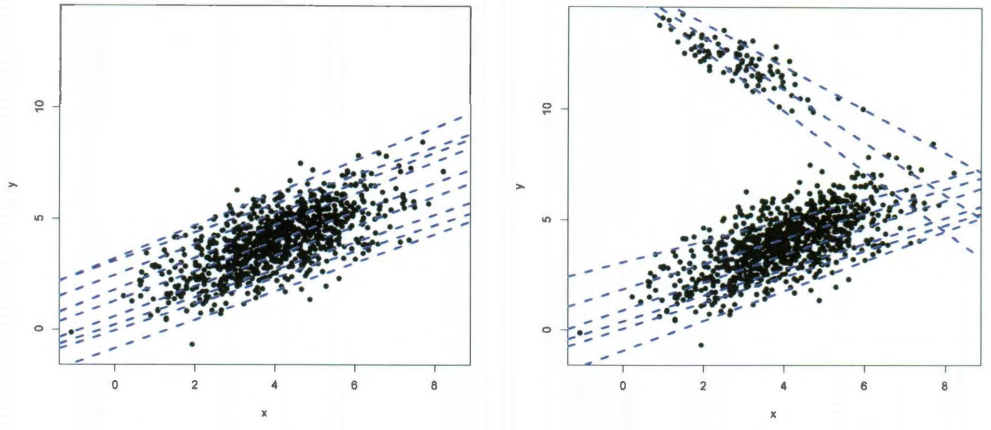
Although quantile regression, like L_1 regression, is considered more robust than least-squares regression, in that it is less affected by outliers, large numbers of outliers can cause large changes in the quantile estimation. In Figure 1.2(b), a cluster of 100 outlier points are added to the previous multivariate data as contamination. As we can see, when performing quantile regression for values of $\tau \geq .90$, the regression line is now not only not passing through the original data, it is trending in the opposite direction of the original data.

1.1.2 Relationship to Maximum Likelihood

If we take $a, b > 0$, we can reparameterize the KB criterion function by

$$\rho_{a,b}(x) = \begin{cases} -ax & \text{if } x < 0 \\ bx & \text{if } x \geq 0. \end{cases} \tag{1.4}$$

It can be shown that this gives an equivalent minimization problem to the minimization in Equation (1.2), where $\tau = \frac{b}{a+b}$. We introduce another function by



(a) Without Contamination (b) With Contamination

Figure 1.2 : Standard linear quantile regression. The .01, .05, .1, .25, .5, .75, .9, .95, and .99 quantile levels are shown.

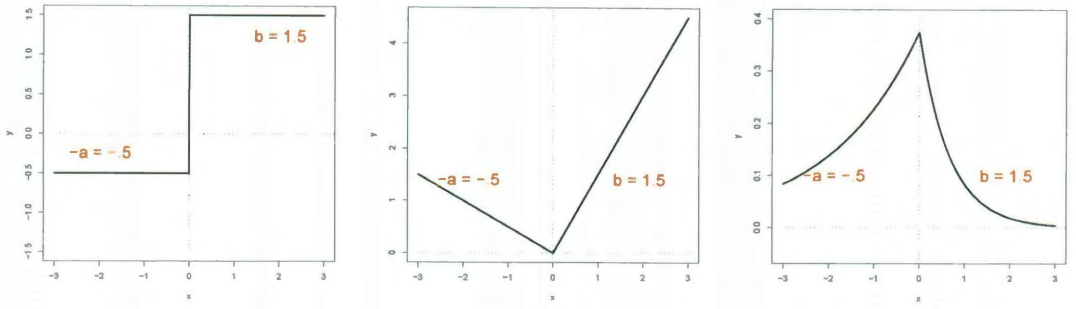
$$g_{a,b}(x) = \frac{d}{dx}\rho_{a,b}(x) = \begin{cases} -a & \text{if } x < 0 \\ b & \text{if } x \geq 0. \end{cases} \tag{1.5}$$

Note that $\rho_{a,b} = x * g_{a,b}(x)$, meaning that $g_{a,b}(x)$ can be thought of as the constant multiplier associated with the value x. We can also create a double exponential distribution by

$$f_{a,b}(x) = c * e^{-\rho_{a,b}(x)}, \tag{1.6}$$

where $c = \frac{ab}{a+b}$, so that the function integrate to 1. The relationships among these three functions are illustrated in Figure 1.3.

We can see that minimizing KB's criterion function is equivalent to fitting a double exponential distribution to data using maximum likelihood estimation (MLE); that



(a) $g_{a,b}(x)$ (b) $\rho_{a,b}(x)$ (c) $f_{a,b}(x)$

Figure 1.3 : Relationship of functions with $a = .5$ and $b = 1.5$.

is,

$$\begin{aligned}\arg\min_{\theta} \sum_{i=1}^{n} \rho_{a,b}(x_i - \theta) &= \arg\max_{\theta} \left(- \sum_{i=1}^{n} \rho_{a,b}(x_i - \theta) \right) \\ &= \arg\max_{\theta} \left(e^{- \sum_{i=1}^{n} \rho_{a,b}(x_i - \theta)} \right) \\ &= \arg\max_{\theta} \prod_{i=1}^{n} f_{a,b}(x_i - \theta).\end{aligned}$$

As with the choice of L_1 error, this form of quantile regression does not have an analytic solution for the MLE, as the derivative of $\rho_{a,b}(x)$ is discontinuous, as seen in Figure 1.3(a). Because of this, more complex methods are used to solve this minimization problem. To solve the problem efficiently, Koenker (1987) uses a modified version of the Simplex algorithm. In particular, he uses a modified version of the algorithm presented by Barrodale and Robert (1973) for efficient L_1 linear approximation.

In an attempt to find an analytic solution for the MLE, a smooth version of the

$g_{a,b}(x)$ function is created, making the derivative of the $\rho_{a,b}(x)$ function analytic. Doing so allows us to use alternative optimization techniques, such as quasi-Newton algorithms, to solve the quantile problem.To create this smooth $g_{a,b}(x)$, "S-curves", such as the cdf of a normal distribution or the cdf of a logistic distribution, are possible options. These S-curves are chosen due to their symmetry as well as their asymptotic nature as $x \to \pm\infty$.

In particular, we look at the cdf of a logistic distribution. We scale the function by $(a+b)$, then shift the function both horizontally and vertically. This makes the asymptote as $x \to -\infty = -a$, the asymptote as $x \to \infty = b$, and makes the function pass through the origin. The general form of this S-curve is

$$g_{a,b,c} = \frac{a+b}{1+exp\{\frac{-(a+b)*c*x}{ab} + log(\frac{b}{a})\}} - a, \tag{1.7}$$

where c is a new tuning parameter that defines the slope of $g_{a,b,c}$ as it passes through the origin. Thus, the greater the value of c, the steeper the slope at the origin.

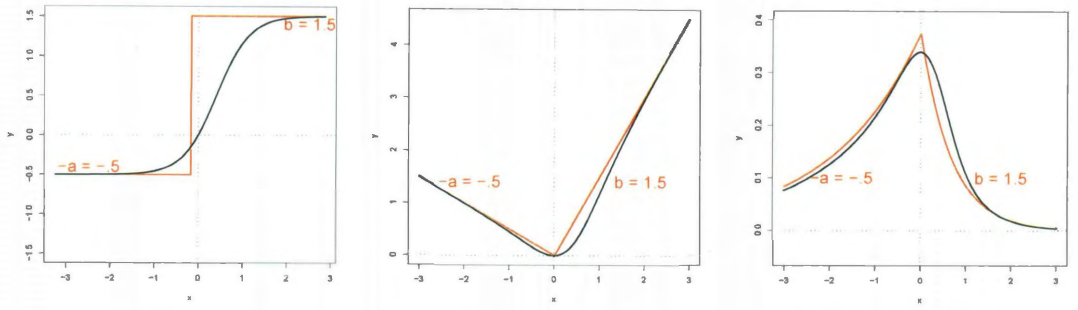
From this new $g_{a,b,c}$ function, we can build smooth versions of both the $\rho_{a,b}$ and $f_{a,b}$ functions by

$$\rho_{a,b,c}(x) = x * g_{a,b,c}(x) \tag{1.8}$$

and

$$f_{a,b,c}(x) = k * e^{-\rho_{a,b,c}(x)}, \tag{1.9}$$

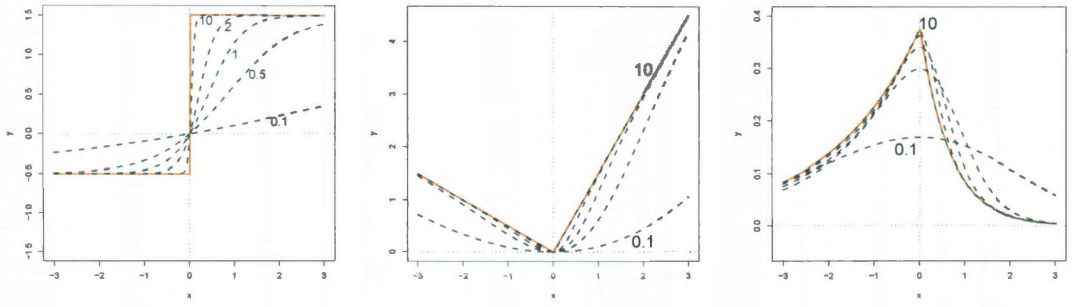
where k is the normalizing constant that makes $\int_{-\infty}^{\infty} f_{a,b,c}(x)dx = 1$. The relationships of these new functions can be seen in Figure 1.4. For comparison, the KB versions of these functions are displayed in red. In Figure 1.5, the resulting functions are plotted for $c = (.1, .5, 1, 2, 10)$. We can see that as $c \to \infty$, $g_{a,b,c} \to g_{a,b}$. Thus, $\rho_{a,b,c} \to \rho_{a,b}$



(a) $g_{a,b,c}(x)$ (b) $\rho_{a,b,c}(x)$ (c) $f_{a,b,c}(x)$

Figure 1.4 : Relationship of smooth functions with $a = .5$, $b = 1.5$, and $c = 1$. The functions from Figure 1.3 are shown in red.

and $f_{a,b,c} \rightarrow f_{a,b}$ as well. In a manner similar to solving the KB criterion function



(a) $g_{a,b,c}(x)$ (b) $\rho_{a,b,c}(x)$ (c) $f_{a,b,c}(x)$

Figure 1.5 : Smooth functions with $a = .5$, $b = 1.5$, and c values of .1, .5, 1, 2, and 10. As the c values increase, the closer the smooth functions resemble the functions from Figure 1.3, shown in red.

for sample quantiles, in particular, solving the equivalent MLE expression, we can fit the smooth double exponential function, $f_{a,b,c}$ to data to obtain quantile estimates.

That is, we can solve the maximization

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} f_{a,b,c}(x_i - \theta). \tag{1.10}$$

However, which quantiles are estimated are no longer determined solely by the a and b parameters in the double exponential. Instead, not only do those parameters matter, the parameter c affects the estimate as well as the type of data itself, making this method parametric. These effects can be seen in Figure 1.6. As we can see, when c is large, the value of $\hat{\theta}$ goes to the sample quantile, the same value the KB criterion function estimates. However, when c is smaller, it tends to bring the estimate closer to the median of the data.

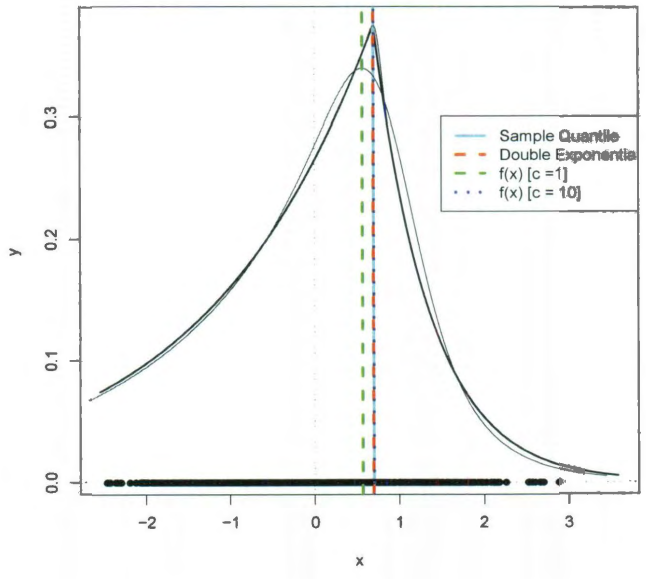


Figure 1.6 : MLE fits of standard double exponential and smooth double exponentials on $N(0,1)$ data using $a = .5$ and $b = 1.5$. The .75 sample quantile, as well as the MLE fit of the standard double exponential distribution, is 0.699. The MLE fit of the smooth double exponential with $c = 1$ is 0.566 while the fit with $c = 10$ is 0.700.

With this added complexity and parametric assumptions comes one advantage: theoretic quantiles estimated can now be determined using common analytic methods. In particular, this can be done maximizing the expectation. For example, if the data come from a $N(0,1)$ distribution, we can find the theoretic quantile achieved, θ_t, as follows:

$$\theta_t = \arg\max_{\theta} E[\log(f_{a,b,c}(x-\theta))] = \arg\min_{\theta} \int_{-\infty}^{\infty} \rho_{a,b,c}(x-\theta) * \phi(x)dx, \qquad (1.11)$$

where $\phi(x)$ is the pdf of a N(0,1) distribution. By using known values of a, b, and c and an assumed distribution for the data, we can derive theoretic quantile levels that the maximum likelihood will estimate. In Figure 1.7, contour maps of these estimated quantile values can be seen for $N(0,1)$ data with a $c = 1$.

Estimation using maximum likelihood with the smooth double exponential also does not add any robustness to the estimation. Just as a large number of outliers will affect the estimation using the KB criterion function, using the smooth version will be similarly affected. To increase robustness, we must turn to a different method.

1.2 Density Estimation with L_2E

L_2 estimation, or L_2E, was developed by Scott (2001) as a robust, parametric density estimator. It belongs to a family of estimators, introduced by Basu et al (1998), but with special computational attractions. To estimate a density $g(x)$ from a sample $(x_1, x_2, \ldots, x_n)$ by a family of distributions $f(x;\theta)$, we find the value of θ by minimizing a data-based estimator of integrated squared error. To see this, we consider

$$\arg\min_{\theta} \int (f(x;\theta) - g(x))^2 \mathrm{d}x, \qquad (1.12)$$

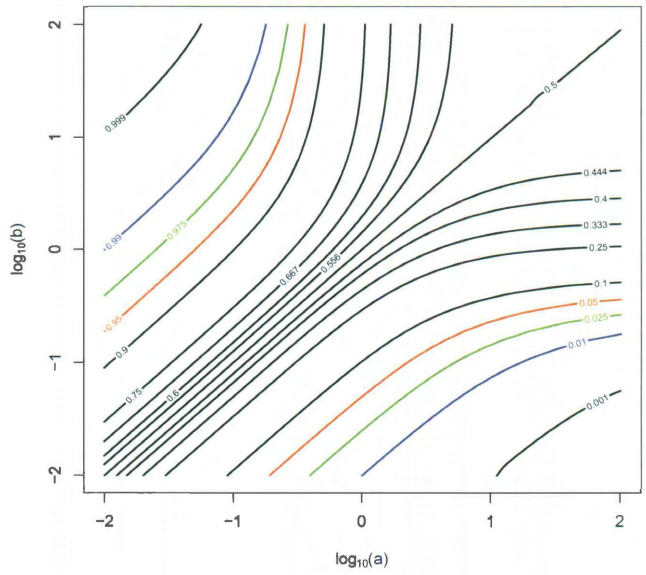


Figure 1.7 : Theoretic MLE quantiles for $N(0,1)$ data with $c = 1$

which expands to

$$\arg\min_{\theta} \int f(x;\theta)^2 \mathrm{d}x - 2\int f(x;\theta)g(x)\mathrm{d}x + \int g(x)^2 \mathrm{d}x. \tag{1.13}$$

Because $g(x)$ doesn't depend on θ, this minimization is equivalent to

$$\arg\min_{\theta} \int f(x;\theta)^2 \mathrm{d}x - 2\int f(x;\theta)g(x)\mathrm{d}x. \tag{1.14}$$

The first term can be computed explicitly, as $f(x;\theta)$ is a known distribution, while the second term is equivalent to $-2E[f(x;\theta)]$, which can be estimated using the sample.

Thus, our L_2E criterion can be estimated in a fully data-based fashion by

$$\arg\min_{\theta} \int f(x;\theta)^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f(x_i;\theta). \tag{1.15}$$

For example, if we believe data to be from a normal distribution, $\phi(x_i;\mu,\sigma)$, we estimate the mean and variance parameters of the normal density, based on a sample $(x_1, x_2, \ldots, x_n)$, by minimizing the quantity

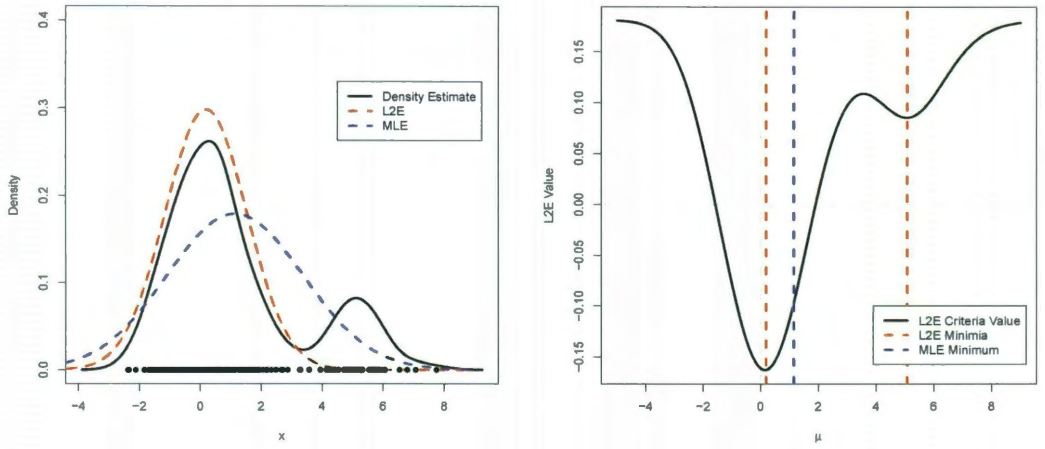
$$(\hat{\mu}_{L_2E}, \hat{\sigma}_{L_2E}) = \arg\min_{\mu,\sigma} \int \phi(x;\mu,\sigma)^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \phi(x_i;\mu,\sigma), \tag{1.16}$$

which, after analyzing the integral, becomes

$$(\hat{\mu}_{L_2E}, \hat{\sigma}_{L_2E}) = \arg\min_{\mu,\sigma} \frac{1}{2\sqrt{\pi}\sigma} - \frac{2}{n}\sum_{i=1}^{n} \phi(x_i;\mu,\sigma). \tag{1.17}$$

To illustrate this, we take a sample $X = (x_1, x_2, \ldots, x_{250})$ with 200 points from a $N(0,1)$ distribution, which we consider our uncontaminated data, and 50 points from a $N(5,1)$ distribution, which we consider contamination. By minimizing the quantity in equation (1.17), we obtain estimates of $\hat{\mu}_{L_2E} = 0.1626$ and $\hat{\sigma}_{L_2E} = 1.3380$. However, if we use maximum likelihood to estimate the parameters, we obtain $\hat{\mu}_{MLE} = 1.1274$ and $\hat{\sigma}_{MLE} = 2.2266$. This is compared to the sample mean and standard deviation of the uncontaminated data, $\hat{\mu}_{samp} = 0.1452$ and $\hat{\sigma}_{samp} = 1.0663$. A comparison of the estimated density functions, along with a kernel density estimate of the data, can be seen in Figure 1.8(a).

Unlike MLE, the L_2E equation is not convex. This is apparent in Figure 1.8(b), which shows a plot of the resulting values of the L_2E equation for a range of μ values and a known $\sigma = 1$. Two distinct local minima can be seen, in particular at $\mu = 0.1528$



(a) (b)

Figure 1.8 : (a) Comparison of L_2E and MLE fits of 200 points of N(0,1) data and 50 points of $N(5,1)$ contamination data, with a kernel density estimate. (b) L_2E criterion values for various values of μ and a known value of $\sigma = 1$. For comparison, local minima of the criterion and $\hat{\mu}_{MLE} = 1.1273$ are marked.

and $\mu = 5.0609$. We note that the global minimum estimates the mean parameter for the distribution of the uncontaminated data while the other local minimum estimates the mean parameter for the distribution of the contaminated data. Because of this, caution must be used exploring initial values in the optimization.

1.2.1 L_2E Linear Regression

Under the assumption of $N(0, \sigma^2)$ residuals, we can adapt equation (1.17) to perform linear regression to estimate the conditional mean of Y given $X = x$. By minimizing the quantity

$$(\hat{\beta}_{L_2E}, \hat{\sigma}_{L_2E}) = \arg\min_{\beta,\sigma} \frac{1}{2\sqrt{\pi}\sigma} - \frac{2}{n}\sum_{i=1}^{n} \phi(y_i - x_i'\beta; 0, \sigma), \tag{1.18}$$

we obtain both the coefficients of the linear regression line as well as an estimate for the standard deviation of the residuals. An example of this type of regression can be seen in Figure 1.9, compared to least squares linear regression on both the full data set and the uncontaminated data. The data come from 900 points of multivariate normal data centered at $(4, 4)$, considered to be the uncontaminated data, combined with 100 points of multivariate data centered at $(3, 10)$, considered to be contamination data. As we can see, the L_2E regression line computed from the full data set is a good approximation of the least squares regression line computed from only the uncontaminated data.

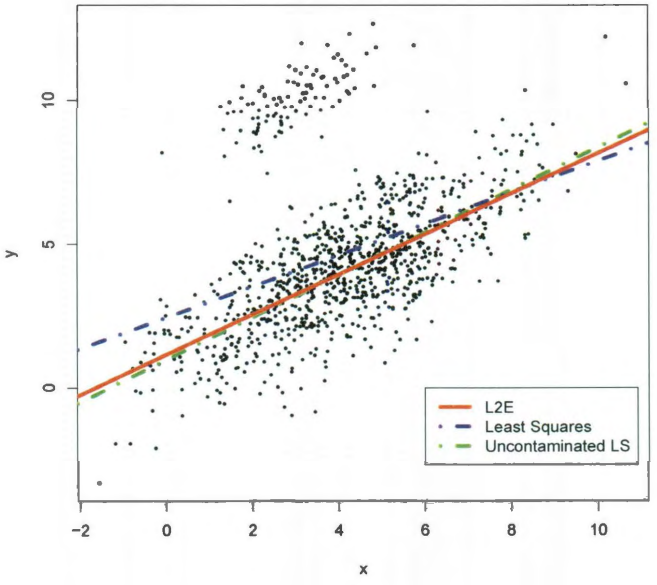


Figure 1.9 : L_2E regression compared with least squares regression (for both full and uncontaminated data) on multivariate normal data with contamination.

1.3 Discussion

There are some situations where KB's quantile regression, that is, the generalization of L_1 regression, can suffer issues with robustness. Therefore, in order to improve upon the robustness of quantile regression, L_2E methods were examined for robust mean regression to determine a similar way to adapt KB's quantile regression. Chapter 2 describes the resulting L_2E method that adds robustness to conditional quantile estimation. The theory and asymptotic behavior of this method are discussed in Chapter 3. Nonlinear and semi-parametric applications are described in Chapter 4. Examples on simulated data with many outliers are shown in Chapter 5 while examples on real data are shown in Chapter 6.

Chapter 2

Robust Quantile Regression

As we have shown, L_2E is able to be used as a robust density estimator and as a robust criteria for regression. From this, it is natural for us to believe that we can somehow use L_2E as a robust quantile estimator and, from that, a robust criteria for quantile regression. From this, we can develop methods for both estimating the coefficients of a robust quantile regression model as well as the variances of those coefficients and a criteria to measure the fit of that model.

2.1 Estimating Quantiles with L_2E

Just as fitting a double exponential distribution to data using maximum likelihood obtains sample quantiles, we can obtain sample quantiles in a similar manner using L_2E. In Equation (1.15), if we take $f(x;\theta)$ to be the double exponential function, that is

$$f(x;\theta) = f_{a,b}(x-\theta) = \begin{cases} \frac{ab}{a+b}e^{a(x-\theta)} & \text{if } x < \theta \\ \frac{ab}{a+b}e^{-b(x-\theta)} & \text{if } x \geq \theta, \end{cases} \tag{2.1}$$

we can minimize the quantity to estimate quantiles from a random sample from a known density function. The values of a and b affect the quantile estimated, as does the true distribution, $g(x)$, making this method parametric. Our L_2E minimization

estimate becomes

$$\arg\min_{\theta} \int f_{a,b}(x-\theta)^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f_{a,b}(x_i-\theta). \tag{2.2}$$

Given a distribution $g(x)$, we can determine theoretic quantiles estimated by using L_2E with a double exponential function given values of a and b. Using our known distributions $g(x)$ and $f(x;\theta)$, we can evaluate the minimization in equation (1.14). In fact, because $f(x;\theta) = f(x-\theta)$, the value of the first term in the equation doesn't depend on θ, reducing the minimization to

$$\arg\min_{\theta} \ -2\int f(x;\theta)g(x)\mathrm{d}x. \tag{2.3}$$

Unlike the MLE, where the theoretic quantile level is known to be $\frac{b}{a+b}$, the values of a and b that achieve desired quantile levels vary depending on the distribution $g(x)$. As seen in Section 3.1, it is possible for the theoretic quantile level to be $\frac{b}{a+b}$, such as in the case where $g(x) \sim \text{Unif}(0,1)$, but that does not hold for all distributions. For example, if we wanted the .75 quantile when $g(x) \sim N(0,1)$, and we apply the constraint $a+b=2$, we would use the values $a = 0.382$ and $b = 1.618$.

In Figure 2.1, a double exponential distribution with $a = 0.382$ and $b = 1.618$ are used to obtain the L_2E estimate of the .75 quantile from 900 points of $N(0,1)$ data. As we can see, the estimated value of 0.6381 is very close to the sample quantile, 0.6101. In Figure 2.2(a), we can see that if we add 100 points of $N(5,.1)$ contamination data to our sample, we still obtain a close estimate of the .75 quantile of the uncontaminated data. In the plot, the red dot marks the .75 quantile for the full data set, which is the maximum likelihood estimate. The added robustness of L_2E is particularly apparent in fig 2.2(b), where the .90 quantile is estimated. Once again, the L_2E estimate is

very close to the sample .90 quantile of the uncontaminated data, while the maximum likelihood estimate now appears in the contamination data.

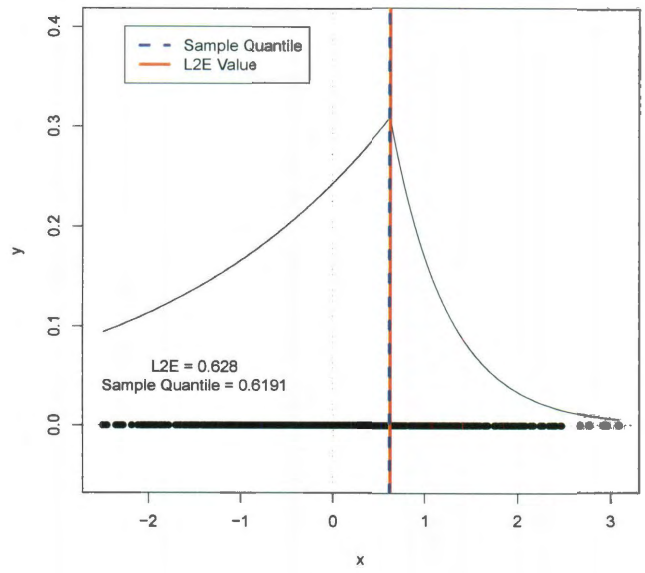
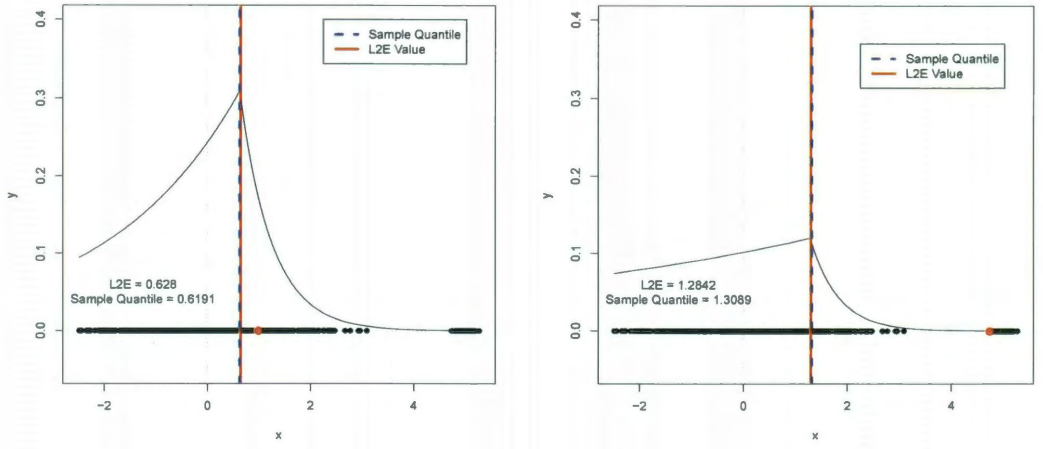


Figure 2.1 : L_2E estimate, using $a = 0.382$ and $b = 1.618$, of the .75 quantile of 900 points of $N(0, 1)$. The sample .75 quantile of the data is also marked.

As before with the MLE, the smooth version of the double exponential function from equation (1.9) can be used in hopes of finding an analytic solution. Given values of a, b, and c, as well as a distribution $g(x)$, we can determine theoretic quantiles in the same manner as before using equation (2.3). Although the theoretic quantile levels are different than with the regular double exponential function using the same values of a and b, the estimation works in a similar manner, maintaining the added robustness.



(a) .75 Quantile

(b) .90 Quantile

Figure 2.2 : L_2E estimates of the .75 and .90 quantiles of 900 points of $N(0,1)$ with 100 points of $N(5,.1^2)$ contamination. The sample quantiles of the uncontaminated data are also marked. The red dot in the data represents the maximum likelihood estimates of the quantiles.

2.2 L_2E Quantile Regression

Just as before in Section 1.2.1, we can adapt our L_2E minimization to perform quantile regression. In particular, by adapting the minimization in Equation (2.2) to be

$$\arg\min_{\beta} \int f_{a,b}(x)^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f_{a,b}(y_i - x_i'\beta), \tag{2.4}$$

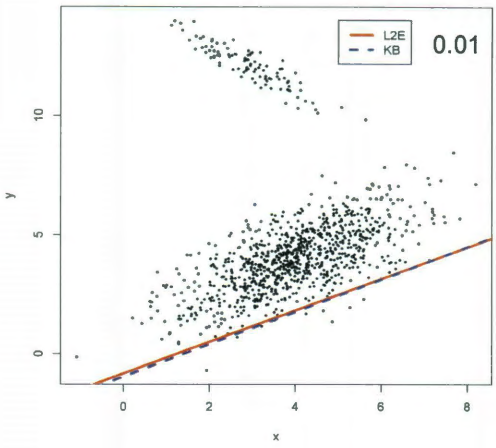
we can obtain estimates of the coefficients of the linear quantile regression equation. Because L_2E quantile estimation is parametric, an assumption about the residuals must be made.

For example, if we assume that the residuals are distributed $N(0,1)$, we can use the same method as before to determine which values of a and b should be used in order to estimate the desired quantile level. So, if the target quantile to be estimated

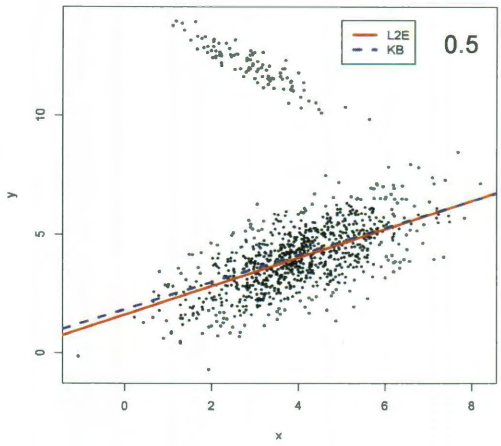
was the .75 quantile, setting $a = 0.382$ and $b = 1.618$ would give us the desired coefficient estimates.

Figure 2.3 shows a comparison of L_2E quantile regression with KB's quantile regression. The data set includes 900 points of bivariate normal data, generated such that the residuals around the mean line are approximated distributed $N(0, 1)$, and 100 points of contamination data placed above the cloud of normal data. As we can see, the regression lines are similar for both the .01 quantile level and the .50 quantile level. However, once the desired quantile level goes above .90, the regression line from KB's method jumps up to the contamination cloud. As we can see in Figure 2.3(c), not only does the KB line pass through the contamination cloud, it trends the opposite direction from the uncontaminated data. However, the L_2E regression line remains in the uncontaminated cloud, still providing an estimate of the .99 quantile of the non-contamination data. In Figure 2.4, we see a comparison of L_2E quantile regression with KB's quantile regression on the full data set, as well as Koenker's quantile regression on the non-contaminated data.

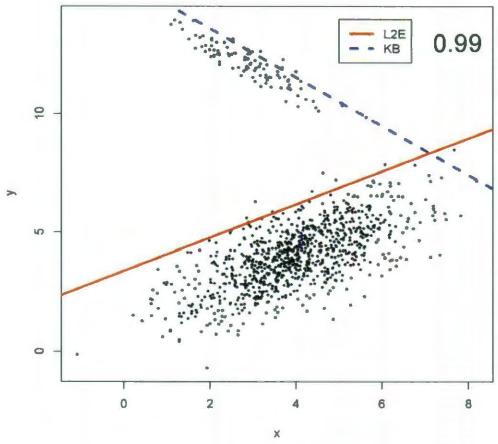
Although we might be able to assume $N(0, \sigma^2)$ residuals about the mean residual line, assuming $N(0, 1)$ is a stretch. However, if we are able to estimate σ in a robust fashion first, we can still perform quantile regression. One such way to estimate σ is to perform the L_2E linear regression outlined in Section 1.2.1. This gives us a robust estimate of σ which can then be used to either determine values of a and b by solving the minimization problem in Equation 2.3 to obtain the desired quantile, or to scale the data so that the residuals are distributed $N(0, 1)$. To perform the latter option, we first obtain our estimate $\hat{\sigma}_{L_2E}$, scale the data by dividing our response variable by $\hat{\sigma}_{L_2E}$, perform L_2E quantile regression as we did before with our $N(0, 1)$ residuals, and then rescaling the parameters by $\hat{\sigma}_{L_2E}$.



(a) .01 quantile regression



(b) .50 quantile regression



(c) .99 quantile regression

Figure 2.3 : Comparison of L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on 900 points of bivariate normal data with 100 points of contamination added above. Least squares residuals are assumed to be $N(0, 1)$

Once again, we can replace the double exponential function, $f_{a,b}(x)$, in Equation 2.2 with the smooth double exponential function $f_{a,b,c}(x)$ to achieve similar results. By doing so, the stability of the optimization, and thus estimation, increases. This

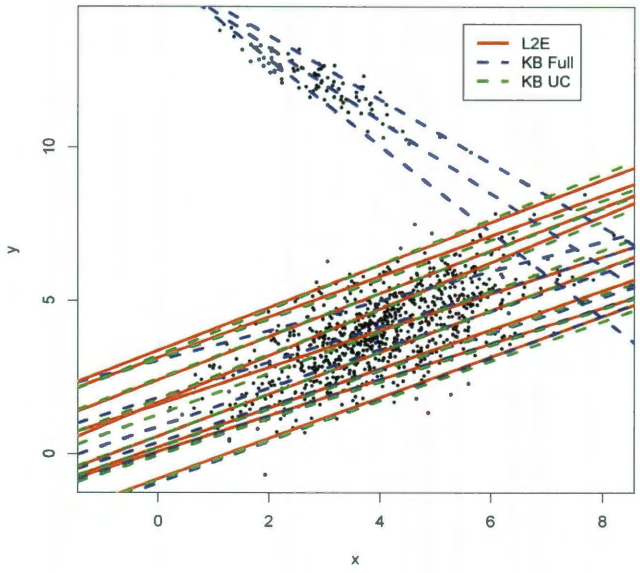
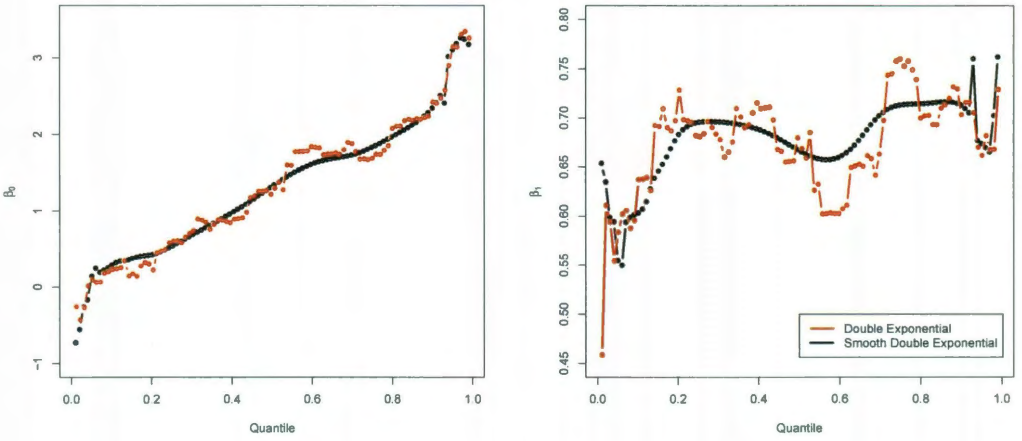


Figure 2.4 : Comparison of L_2E quantile regression, shown in red, KB's quantile regression on the full data set, shown in blue, and KB's quantile regression on the uncontaminated data on 900 points of bivariate normal data with 100 points of contamination added above. Least squares residuals are assumed to be $N(0, 1)$

can be seen in Figure 2.5, where the x-axis represents the desired quantile level and the y-axis represents the estimated value of the intercept and slope coefficients from L_2E quantile regression for those desired quantile levels. These plots allow us to see the different effects that each predictor variable across different quantile levels. As we can see, the coefficient estimates from using $f_{a,b}(x)$, shown in red, are noticeably less stable, as they bounce around the estimates from using $f_{a,b,c}(x)$, shown in black. Because of this property, the smooth double exponential distribution will be used for L_2E quantile regression in future examples unless otherwise noted.



(a) Intercept (b) Slope

Figure 2.5 : Comparison of the estimated slope and intercept coefficients for L_2E quantile regression using the same data found in Figure 2.4. L_2E coefficient estimates using the double exponential function, $f_{a,b}(x)$, are shown in red, while the L_2E coefficient estimates using the smooth double exponential function, $f_{a,b,c}(x)$ are shown in black.

2.2.1 Estimating Regression Coefficient Variances

As shown in Section 3.2.3, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_{L_2E} - \beta_*) \to MN\left(\tilde{0}_{p+1}, \frac{E\left[\left(f'_{a,b,c}(\epsilon)\right)^2\right]}{E\left[f''_{a,b,c}(\epsilon)\right]^2} E[(XX')]^{-1}\right),$$

where $\hat{\beta}_{L_2E}$ is the vector containing the quantile regression coefficients, β_* is the vector of true coefficients, $f'_{a,b,c}(s) = \frac{\partial}{\partial s} f_{a,b,c}(s)$, and $f''_{a,b,c}(s) = \frac{\partial^2}{\partial s^2} f_{a,b,c}(s)$. From this can estimate the covariance matrix of $\hat{\beta}_{L_2E}$ to be

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(f'_{a,b,c}(y_i - x_i'\beta))^2}{\left(\frac{1}{n}\sum_{i=1}^{n} f''_{a,b,c}(y_i - x_i'\beta)\right)^2}[M'M]^{-1},$$

where M is the $n \times (p+1)$ data matrix. We can use this estimated covariance matrix to estimate the variance of each coefficient. This allows us to create confidence intervals and p-values for each coefficient.

2.2.2 Model Selection Using AIC

As our solution to the minimization found in Equation 2.4 occurs when

$$\sum_{i=1}^{n} \nabla_{\beta} f_{a,b,c}(y_i - x_i'\beta) = \tilde{0}_{p+1},$$

we can treat our L_2E quantile regression criteria as an M-estimator. Thus, we can create a robust Akaike Information Criterion (AIC) using the results found in Ronchetti (1997). That is, we can define the AIC for a L_2E quantile regression model by

$$2\sum_{i=1}^{n}\left(\int f_{a,b,c}(x)^2 \mathrm{d}x - 2f_{a,b,c}(y_i - x_i'\hat{\beta})\right) + \alpha_p,$$

where $\hat{\beta}$ is the vector of estimated L_2E quantile regression coefficients and $\alpha_p = 2\ tr(S^{-1}Q)$. Because

$$\begin{aligned} S &= -E[H_{\beta} f_{a,b,c}(y_i - x_i'\beta)] \\ &= -E\left[f''_{a,b,c}(Y - X'\beta)(XX')\right] \\ &= -E\left[f''_{a,b,c}(Y - X'\beta)\right] E\left[(XX')\right] \end{aligned}$$

and

$$\begin{aligned} Q &= E[(\bigtriangledown_\beta f_{a,b,c}(y_i - x_i'\beta))] \\ &= E\left[\left(f'_{a,b,c}(Y - X'\beta)X\right)\left(f'_{a,b,c}(Y - X'\beta)X\right)'\right] \\ &= E\left[\left(\left(f'_{a,b,c}(Y - X'\beta)\right)^2 (XX')\right)\right] \\ &= E\left[\left(f'_{a,b,c}(Y - X'\beta)\right)^2\right] E\left[(XX')\right], \end{aligned}$$

α_p reduces to

$$\begin{aligned} \alpha_p &= -2\frac{E\left[\left(f'_{a,b,c}(Y - X'\beta)\right)^2\right]}{E[\left[f''_{a,b,c}(Y - X'\beta)\right]} tr(E\left[(XX')\right]^{-1} E\left[(XX')\right]) \\ &= -2\frac{E\left[\left(f'_{a,b,c}(\epsilon)\right)^2\right]}{E[\left[f''_{a,b,c}(\epsilon)\right]} tr(I_{p+1}) \\ &= -2\frac{E\left[\left(f'_{a,b,c}(\epsilon)\right)^2\right]}{E[\left[f''_{a,b,c}(\epsilon)\right]} (p+1), \end{aligned}$$

where ϵ has the assumed distribution of the residuals. This criterion acts in the same way as the standard AIC, in that models with lower AIC values are considered to be better fits to the data. Like Koenker (2005), in practice we use the median regression criterion to determine the model with the best fit for all quantile lines, although it may be possible to use this criterion with other quantile regression levels. We also note that in the implementation of our algorithm, the residuals are scaled for the model fit, so care must be taken as the scaling does have an effect on our criterion.

We believe that other information criteria developed for M-estimators, such as the Schwarz Information Criteria (SIC) described in Machado (1993), can be used to create additional model selection criteria for L_2E quantile regression. For example,

the SIC can be shown to be

$$\sum_{i=1}^{n}\left(\int f_{a,b,c}(x)^2 \mathrm{d}x - 2f_{a,b,c}(y_i - x_i'\hat{\beta})\right) + \frac{1}{2}(p+1)log(n),$$

where p is the number of factors in the model. In initial testing, this SIC behaves very well on simulated data. However, it does not appear to behave as well as the AIC on real data. We postpone a detailed evaluation and comparison of AIC and SIC.

2.3 Discussion

In order to develop a robust method of estimating conditional quantiles, we turn to L_2E density estimation and adapt it to estimate quantiles by taking ideas from KB's quantile estimation method. In doing so, we created a robust criteria, and an algorithm to use that criteria, that can be used to perform L_2E quantile regression, giving us robust coefficients for linear models. Then, using methods developed for M-estimators, we are able to find estimates for variances of the regression coefficients as well as a robust version of AIC to assess model selection. The implementation of these ideas have been written in R, the function descriptions of which can be found in Appendix A.

Chapter 3

Theoretic Results

Having discussed how it is possible to perform robust quantile estimation using L_2E, it is important to know how to select parameters for our estimator to achieve specific quantiles and how the selection of those parameters affect the accuracy of the estimator. To do so, in this chapter we examine both the theoretic results of our L_2E quantile estimator as well as its asymptotic behavior.

3.1 Theoretic Values

When using L_2E quantile estimation, it is of particular interest to know what values of a and b in our double exponential achieve specific quantiles. We recall that when we perform quantile estimation using KB's check function criteria, values of a and b achieve the $\frac{b}{a+b}$th quantile, regardless of distribution. L_2E quantile regression, however, requires a knowledge of the underlying distribution to determine which quantile is achieved by the values of a and b. That is, given a sample $(x_1, x_2, \ldots, x_n)$ from a population X with cdf G(x) and pdf g(x), can find the value θ_{L_2E} that solves the minimization

$$\theta_{L_2E} = \arg\min_{\theta} \int f_{a,b,c}(x;\theta)^2 \mathrm{d}x - 2\int f_{a,b,c}(x;\theta)g(x)\mathrm{d}x.$$

Note that because the first term does not depend on θ, as θ is a shift parameter, it is a constant with regards to θ. This allows us to reduce the minimization to

$$\theta_{L_2E} = \arg\min_{\theta} \ -2\int f_{a,b,c}(x;\theta)g(x)\mathrm{d}x. \tag{3.1}$$

After finding θ_{L_2E} in Equation 3.1, the theoretic quantile level estimated can be found by

$$\tau_{L_2E} = G(\theta_{L_2E}).$$

If we want to determine values of a and b to achieve a particular quantile level, τ, for a set value of c, we first determine the true quantile of X, denoted by θ_0, by taking $\theta_0 = G^{-1}(\tau)$. There are infinitely many combinations of a and b that can achieve this theoretic quantile, so we impose the restriction $a + b = r$, where r is a specified constant. From this, we find the value for a such that θ_{L_2E} from Equation 3.1 is equal to θ_0. Then, b can then be found as $b = r - a$.

For the following examples, we substitute the standard double exponential distribution, $f_{a,b}(x)$, for the smooth double exponential function, $f_{a,b,c}(x)$, in Equation 3.1. This was done to make the derivations in the examples simpler. However, the same methods apply when the smooth double exponential is used.

3.1.1 Uniform(0,1) Example

For given values of a and b and using a Uniform(0,1) distribution for g(x), we get from Equation 3.1

$$\begin{aligned}\theta_{L_2E} &= \arg\min_{\theta}\left[-2\int f_{a,b}(x;\theta)(1)\mathrm{d}x\right] \\ &= \arg\max_{\theta}\left[\int_0^{\theta} ke^{a(x-\theta)}\mathrm{d}x + \int_{\theta}^{1} ke^{-b(x-\theta)}\mathrm{d}x\right] \\ &= \arg\max_{\theta}\left[k\left(\frac{1}{a} - \frac{1}{a}e^{-a\theta} - \frac{1}{b}e^{-b+b\theta} + \frac{1}{b}\right)\right].\end{aligned}$$

To solve this maximization, we take the derivative of the right hand side with respect to θ, set the resulting equation equal to 0, and then solve for θ. Thus, we get

$$\begin{aligned}k\left[e^{-a\theta} - e^{-b+b\theta}\right] &= 0 \\ e^{-a\theta} &= e^{-b+b\theta} \\ -a\theta &= -b + b\theta \\ a\theta + b\theta &= b \\ \theta_{L_2E} &= \frac{b}{a+b}.\end{aligned}$$

From this, we see that an infinite number of combinations of a and b will achieve the same theoretic quantile value. Thus, to find a unique solution, it is necessary to restrict the values by setting $a+b=r$. Note that this is equivalent to the result found by the method presented by KB, particularly with $a+b=1$. However, as evidenced by the following section, this nice result does not hold for all possible distributions for $g(x)$.

3.1.2 Standard Normal Example

Again for given values of a and b and using a Normal(0,1) distribution for $g(x)$, we get from Equation 3.1

$$\begin{aligned}
\theta_{L_2E} &= \arg\min_{\theta}\left[-2\int f_{a,b}(x;\theta)\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}\mathrm{d}x\right] \\
&= \arg\max_{\theta}\left[\int_0^{\theta} ke^{a(x-\theta)}\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}\mathrm{d}x + \int_{\theta}^{1} ke^{-b(x-\theta)}\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}\mathrm{d}x\right] \\
&= \arg\max_{\theta}\left[ke^{-a\theta+\frac{a^2}{2}}\int_{-\infty}^{\theta}\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-a)^2}{2}}\mathrm{d}x + ke^{b\theta+\frac{b^2}{2}}\int_{\theta}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-(x+b)^2}{2}}\mathrm{d}x\right] \\
&= \arg\max_{\theta}\left[ke^{-a\theta+\frac{a^2}{2}}\Phi(\theta-a) + ke^{b\theta+\frac{b^2}{2}}\left(1-\Phi(\theta+b)\right)\right]
\end{aligned}$$

where $\Phi(x)$ is the cdf of a Normal(0,1) distribution. Taking the derivative with respect to θ, setting the result equal to 0, and solving for θ, we get

$$\begin{aligned}
0 &= -ake^{-a\theta+\frac{a^2}{2}}\Phi(\theta-a) + ke^{-a\theta+\frac{a^2}{2}}\phi(\theta-a) + bke^{b\theta+\frac{b^2}{2}}\left(1-\Phi(\theta+b)\right) \\
&\quad - ke^{b\theta+\frac{b^2}{2}}\phi(\theta+b) \\
0 &= -a\Phi(\theta-a) + \phi(\theta-a) + be^{(a+b)\theta+\frac{b^2-a^2}{2}}\left(1-\Phi(\theta+b)\right) \\
&\quad - e^{(a+b)\theta+\frac{b^2-a^2}{2}}\phi(\theta+b) \\
e^{(a+b)\theta+\frac{b^2-a^2}{2}} &= \frac{-a\Phi(\theta-a)+\phi(\theta-a)}{-b\left(1-\Phi(\theta+b)\right)+\phi(\theta+b)} \\
\theta_{L_2E} &= \frac{1}{a+b}ln\left[\frac{\phi(\theta-a)-a\Phi(\theta-a)}{\phi(\theta+b)-b\left(1-\Phi(\theta+b)\right)}\right] + \frac{a-b}{2}.
\end{aligned}$$

From this point, numerical approximation is necessary to solve for θ_{L_2E}. Figure 3.1(a) show the theoretic quantiles achieved by various combinations of a and b, presented in log_{10} scale. As shown in Figure 3.1(b), by selecting a value of r such that $a+b=r$, we can find a unique combination of a and b such that the resulting theoretic quantile θ_{L_2E} gives $\Phi(\theta_{L_2E}) = \tau$, where τ is the desired quantile level. For the case shown, $\tau = 0.75$ and $r = 2$. This gives the values of $a = 0.382$ and $b = 1.618$. This is different from the result given by KB, as these values of a and b would estimate a quantile level of $\tau = .809$. As noted earlier, the L_2E τ's are closer to the median than the KB τ's.

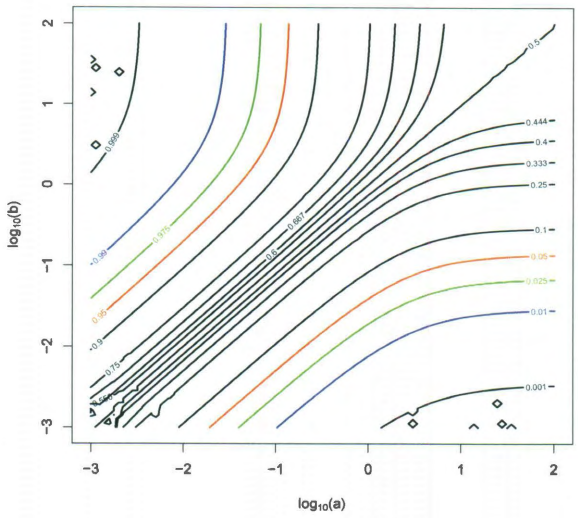
3.1.3 Robust Evaluation Via a Mixture of Uniforms

In order to see the effects of a mixture distribution on our L_2E quantile estimation, we examine the simple case of a uniform mixture. Assume that without loss of generality we allow g(x) to be a mixture of a Uniform(0,1) distribution,with weight $w \in (0,1)$ and a Uniform(u_1,u_2) distribution, with $1 < u_1 < u_2$ and with weight $(1-w)$. Then for given values of a and b, we get from Equation 3.1

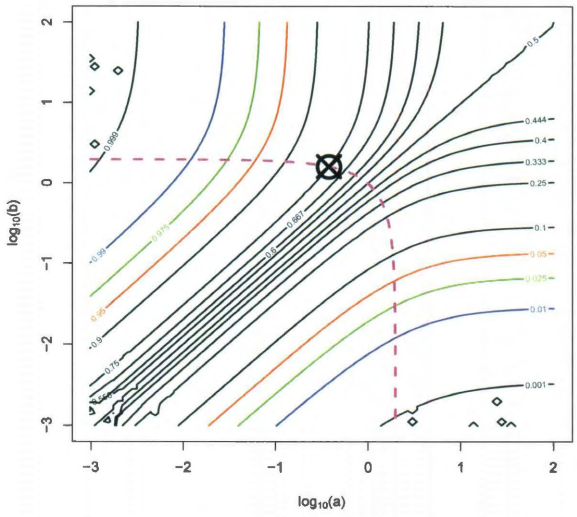
$$\theta_{L_2E} = \arg\min_{\theta} \left[-2w \int f_{a,b}(x;\theta)(1)\mathrm{d}x - 2(1-w) \int f_{a,b}(x;\theta) \frac{1}{u_2 - u_1} \mathrm{d}x \right]. \quad (3.2)$$

Because of the nature of the double exponential distribution, we have to look at several cases. First, we examine the case that $\theta \in (0,1)$, that is, a critical point within the range of the Unif(0,1) distribution. This gives us

$$-2w \int_0^{\theta} k e^{a(x-\theta)} \mathrm{d}x - 2w \int_{\theta}^{1} k e^{-b(x-\theta)} \mathrm{d}x - \frac{2(1-w)}{u_2 - u_1} \int_{u_1}^{u_2} k e^{-b(x-\theta)} \mathrm{d}x.$$



(a) Contours



(b) Contours with $a + b = 2$ trace

Figure 3.1 : Contours of theoretic L_2E quantiles for $g(x) \sim N(0,1)$ for various values of a and b, presented in $\log_{10}$ scale. Plot (b) includes a trace of $a + b = 2$ and marks where that trace crosses the .75 contour line, at $a = 0.382$ and $b = 1.618$

From this, we can see that the critical point in this region, if it exists, can be found by

$$\theta_{L_2E} = \frac{-1}{a+b} ln \left[e^{-b} + \frac{1-w}{w(u_2-u_1)} \left[e^{-bu_2} - e^{-bu_1} \right] \right].$$

From looking at the second derivative, we can see that this point is a local minimum of Equation 3.2. Next, we examine the case that $\theta \in (1, u_1)$, that is, a critical point between the ranges of the two uniform distributions. This gives us

$$-2w \int_0^1 k e^{a(x-\theta)} \mathrm{d}x - \frac{2(1-w)}{u_2-u_1} \int_{u_1}^{u_2} k e^{-b(x-\theta)} \mathrm{d}x.$$

The critical point in this region, if it exists, can be found by

$$\theta = \frac{-1}{a+b} ln \left[\frac{1-w}{w(u_2-u_1)} \frac{\left[e^{-bu_2} - e^{-bu_1} \right]}{e^{a-1}} \right].$$

From the second derivative, we see that this point is a local maximum of Equation 3.2. In the case where $\theta \in (u_1, u_2)$, that is, a critical point within the range of the Unif(u_1,u_2) distribution, we get

$$-2w \int_0^1 k e^{a(x-\theta)} \mathrm{d}x - \frac{2(1-w)}{u_2-u_1} \int_{u_1}^{\theta} k e^{a(x-\theta)} \mathrm{d}x - \frac{2(1-w)}{u_2-u_1} \int_{\theta}^{u_2} k e^{-b(x-\theta)} \mathrm{d}x.$$

The critical point in this region, if it exists, can be found by

$$\theta_{L_2E} = \frac{ln \left[e^{au_1} - \frac{w(u_2-u_1)}{1-w} \left[e^a - 1 \right] \right] + bu_2}{a+b}.$$

Again, from the second derivative, we find that this point is a local minimum of Equation 3.2. For the other two regions, namely $\theta < 0$ and $\theta > u_2$, we can see that no critical points, and thus no local extrema exist. From all of this, we see that it is possible for our L_2E equation to have two local minima.

For example, in the case where we have the mixture

$$\frac{3}{5}\text{Unif}(0,1) + \frac{2}{5}\text{Unif}(3,4),$$

there are two local minima, as exhibited in Figure 3.2(a). However, if have the case where we have the mixture

$$\frac{4}{5}\text{Unif}(0,1) + \frac{1}{5}\text{Unif}(2,3),$$

there is only a single local minimum, as seen in Figure 3.2(b).

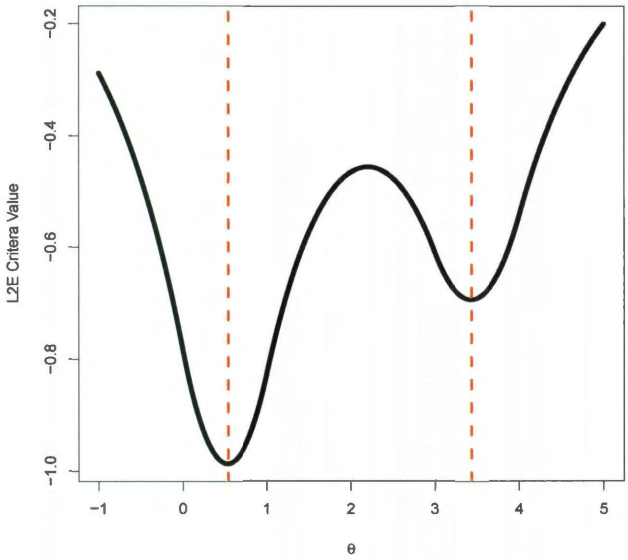
3.2 Asymptotic Theory

In order to determine the asymptotic behavior of of estimate, θ_{L_2E}, we begin once again with the minimization

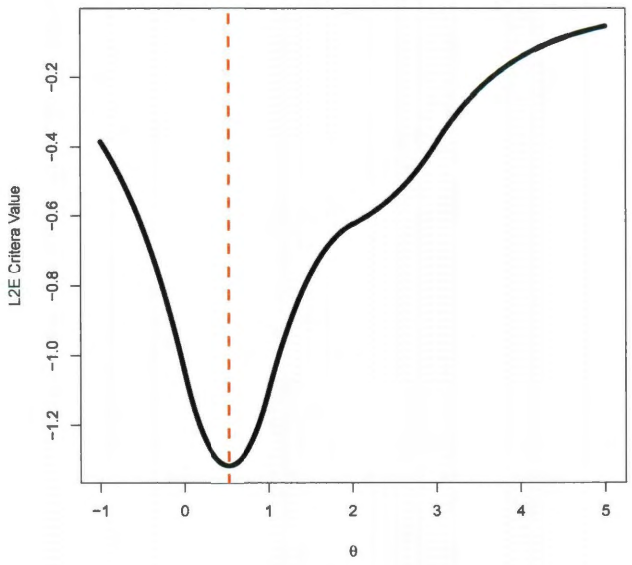
$$\arg\min_{\theta} \int f_{a,b}(x-\theta)^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f_{a,b}(x_i - \theta). \tag{3.3}$$

This has a minimum when the derivative with respect to θ is equal to 0. Again, because the first term does not depend on θ, this is equivalent to finding θ such that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} f_{a,b}(x_i - \theta) = 0. \tag{3.4}$$



(a) $\frac{3}{5}\text{Unif}(0,1) + \frac{2}{5}\text{Unif}(3,4)$



(b) $\frac{4}{5}\text{Unif}(0,1) + \frac{1}{5}\text{Unif}(2,3)$

Figure 3.2 : L_2E criteria values for values of θ with $g(x)$ being a mixture of two uniform distributions. Local minima are marked by the red lines.

This allows us to treat our L_2E estimate as an M-Estimator and use the same methods, such as those described in Van der Vatt (2000), to show asymptotic normality. To do this, we define $\psi_\theta(x_i) = \frac{\partial}{\partial\theta} f_{a,b}(x_i - \theta)$ and $\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(x_i)$. From here, we perform the Taylor expansion of $\Psi_n(\theta)$ about the desired quantile value, that is, θ_0 and find that

$$\begin{aligned} 0 = \Psi_n(\theta_{L_2E}) &= \Psi_n(\theta_0) + (\theta_{L_2E} - \theta_0)\frac{\partial}{\partial\theta}\Psi_n(\theta) + \frac{1}{2}(\theta_{L_2E} - \theta_0)^2\frac{\partial^2}{\partial\theta^2}\Psi_n(\tilde{\theta}_0) \\ (\theta_{L_2E} - \theta_0) &= \frac{-\Psi_n(\theta_0)}{\frac{\partial}{\partial\theta}\Psi_n(\theta_0) + \frac{1}{2}(\theta_{L_2E} - \theta_0)\frac{\partial^2}{\partial\theta^2}\Psi_n(\tilde{\theta}_0)} \\ \implies \sqrt{n}(\theta_{L_2E} - \theta_0) &= \frac{-\sqrt{n}\Psi_n(\theta_0)}{\frac{\partial}{\partial\theta}\Psi_n(\theta_0) + \frac{1}{2}(\theta_{L_2E} - \theta_0)\frac{\partial^2}{\partial\theta^2}\Psi_n(\tilde{\theta}_0)}, \end{aligned}$$

Where $\tilde{\theta}$ is some value between θ_{L_2E} and θ. Examining the right hand side, we first see that $-\sqrt{n}\Psi_n(\theta_0) = -\frac{1}{\sqrt{n}}\Psi_n(\theta_0)$, which by the Central Limit Theorem has a Normal distribution with mean = 0 and a variance that can be found by

$$E_{g(x)}\left[(\frac{\partial}{\partial\theta} f_{a,b}(x_i - \theta_0))^2\right] = \int (\frac{\partial}{\partial\theta} f_{a,b}(x - \theta_0))^2 g(x)\mathrm{d}x.$$

We also see that

$$\frac{\partial}{\partial\theta}\Psi_n(\theta_0) \to_p \frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta} f_{a,b}(x - \theta_0)g(x)\mathrm{d}x$$

and that

$$\frac{1}{2}(\theta_{L_2E} - \theta_0)\frac{\partial^2}{\partial\theta^2}\Psi_n(\tilde{\theta}) \to_p 0.$$

From all of this, and using Slutsky's theorem, we can see that

$$\sqrt{n}(\theta_{L_2E} - \theta_0) \to N\left(0, \frac{\int (\frac{\partial}{\partial\theta} f_{a,b}(x - \theta_0))^2 g(x)\mathrm{d}x}{\left(\frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta} f_{a,b}(x - \theta_0)g(x)\mathrm{d}x\right)^2}\right). \tag{3.5}$$

The same result holds if the smooth double exponential distribution, $f_{a,b,c}(x)$ is used in place of the standard double exponential distribution, $f_{a,b}(x)$.

3.2.1 Uniform(0,1) Example

We take g(x) to be the Uniform(0,1) distribution and values of a and b such that the theoretic value of $\theta_{L_2E} = \theta_0$, where $\theta_0 = \frac{b}{a+b} = \tau$. From this, we see that

$$\begin{aligned}\int \left(\frac{\partial}{\partial\theta} f_{a,b}\left(x-\theta\right)\right)^2 g(x)\mathrm{d}x &= \int_0^\theta \left[a^2k^2e^{2a(x-\theta)}\right]\mathrm{d}x + \int_\theta^1 \left[b^2k^2e^{-2b(x-\theta)}\right]\mathrm{d}x \\ &= \frac{1}{2}k^2\left[a - ae^{-2a\theta}\right] - \frac{1}{2}k^2\left[be^{-2b(1-\theta)} - b\right] \\ &= \frac{1}{2}k^2\left[a + b - ae^{-2a\theta} - be^{-2b(1-\theta)}\right]\end{aligned}$$

and also

$$\begin{aligned}\frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta} f_{a,b}(x-\theta)g(x)\mathrm{d}x &= \frac{\partial}{\partial\theta}\left[\int_0^\theta [-ake^{a(x-\theta)}]\mathrm{d}x + \int_\theta^1 [bke^{-b(x-\theta)}]\mathrm{d}x\right] \\ &= \frac{\partial}{\partial\theta}\left[ke^{-a\theta} - k - ke^{-b(1-\theta)} + k\right] \\ &= -kae^{-a\theta} - kbe^{-b(1-\theta)}.\end{aligned}$$

Therefore, we have

$$\sqrt{n}(\theta_{L_2E} - \theta_0) \to N\left(0, \frac{\left[a + b - ae^{-2a\theta_0} - be^{-2b(1-\theta_0)}\right]}{2(ae^{-a\theta_0} + be^{-b(1-\theta_0)})^2}\right).$$

The theoretic standard deviations for a range of values of a and b can be found in Figure 3.3(a).

3.2.2 Standard Normal Example

If we instead take g(x) to be the standard Normal(0,1) distribution and values of a and b such that the theoretic value of $\theta_{L_2E} = \theta_0$, where $\Phi(\theta_0) = \tau$. We then see that

$$\begin{aligned}
\int (\frac{\partial}{\partial\theta} f_{a,b}(x-\theta))^2 g(x)\mathrm{d}x &= \int_{-\infty}^{\theta} \left[a^2k^2e^{2a(x-\theta)}\right] \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\mathrm{d}x \\
&\quad + \int_{\theta}^{\infty} \left[b^2k^2e^{-2b(x-\theta)}\right] \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\mathrm{d}x \\
&= a^2k^2e^{-2a\theta+2a^2} \int_{-\infty}^{\theta} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-2a)^2}{2}}\mathrm{d}x \\
&\quad + b^2k^2e^{2b\theta+2b^2} \int_{\theta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x+2b)^2}{2}}\mathrm{d}x \\
&= a^2k^2e^{-2a\theta+2a^2}\Phi(\theta-2a) + b^2k^2e^{2b\theta+2b^2}\left[1-\Phi(\theta+2b)\right]
\end{aligned}$$

and also

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta} f_{a,b}(x-\theta)g(x)\mathrm{d}x \\
&= \frac{\partial}{\partial\theta}\left[\int_{-\infty}^{\theta}\left[-ake^{a(x-\theta)}\right]\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}\mathrm{d}x + \int_{\theta}^{\infty}\left[bke^{-b(x-\theta)}\right]\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}\mathrm{d}x\right] \\
&= \frac{\partial}{\partial\theta}\left[-ake^{-a\theta+a^2}\int_{-\infty}^{\theta}\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-a)^2}{2}}\mathrm{d}x + bke^{b\theta+b^2}\int_{\theta}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-(x+b)^2}{2}}\mathrm{d}x\right] \\
&= \frac{\partial}{\partial\theta}\left[-ake^{-a\theta+a^2}\Phi(\theta-a) + bke^{b\theta+b^2}\left[1-\Phi(\theta+b)\right]\right] \\
&= a^2ke^{-a\theta+a^2}\Phi(\theta-a) - ake^{-a\theta+a^2}\phi(\theta-a) \\
&\quad + b^2ke^{b\theta+b^2}\left[1-\Phi(\theta+b)\right] - bke^{b\theta+b^2}\phi(\theta+b).
\end{aligned}$$

This gives us

$$\sqrt{n}(\theta_{L_2E}-\theta_0) \to N\left(0, \frac{a^2e^{-2a\theta_0+2a^2}\Phi(\theta_0-2a)+b^2e^{2b\theta_0+2b^2}[1-\Phi(\theta_0+2b)]}{\left(a^2e^{-a\theta_0+a^2}\Phi(\theta_0-a)-ae^{-a\theta_0+a^2}\phi(\theta_0-a)+b^2e^{b\theta-0+b^2}[1-\Phi(\theta_0+b)]-be^{b\theta_0+b^2}\phi(\theta_0+b)\right)^2}\right).$$

The theoretic standard deviations for a range of values of a and b can be found in Figure 3.3(b).

3.2.3 L_2E Linear Quantile Regression Coefficients

One area of particular interest is the asymptotic behavior of the coefficient estimates of our L_2E quantile regression. We know that given the linear model

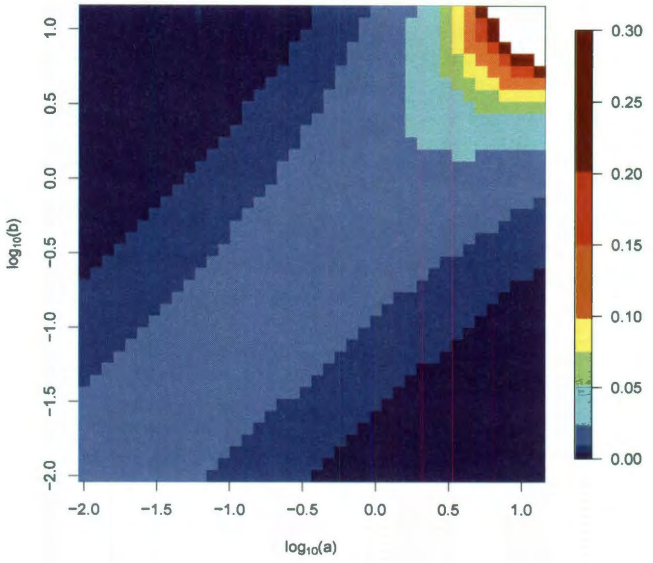
$$Y = X'\beta + \epsilon,$$

the criterion function for L_2E linear quantile regression is

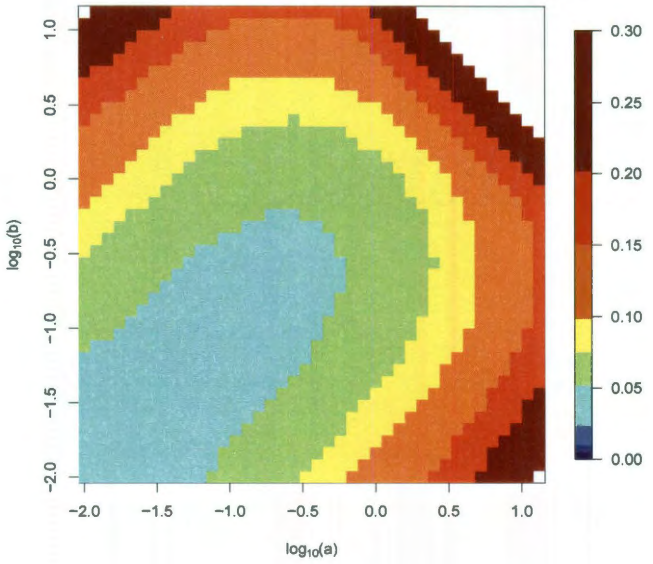
$$\arg\min_{\beta} \int f_{a,b,c}(x)^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f_{a,b,c}(y_i - x_i'\beta),$$

where $f_{a,b,c}$ is the smooth version of the double exponential distribution, y_i are iid from random variable Y, $\beta = \{\beta_0, \beta_1, \ldots, \beta_p\}$ and $x_i = \{1, x_{i1}, \ldots, x_{ip}\}$, which are iid from random variable X, we see that the solution to this minimization occurs when

$$\frac{1}{n}\sum_{i=1}^{n} \bigtriangledown_{\beta} f_{a,b,c}(y_i - x_i'\beta) = \tilde{0}_{p+1}.$$



(a) Uniform



(b) Normal

Figure 3.3 : Theoretic standard deviations for L_2E quantile estimates given various values of a and b and a sample size of 1000.

Let β_* be the value such that $E[\bigtriangledown_\beta f_{a,b,c}(Y - X^{'}\beta_*)] = \tilde{0}_{p+1}$. Then, by performing a Taylor expansion around β_*, we see that

$$\tilde{0}_{p+1} = \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown_\beta f_{a,b,c}(y_i - x_i'\beta_*) + \frac{1}{n}\sum_{i=1}^{n} H_\beta f_{a,b,c}(y_i - x_i'\beta)(\beta_{L_2E} - \beta_*)$$

$$\frac{1}{n}\sum_{i=1}^{n} H_\beta f_{a,b,c}(y_i - x_i'\beta)(\beta_{L_2E} - \beta_*) = -\frac{1}{n}\sum_{i=1}^{n} \bigtriangledown_\beta f_{a,b,c}(y_i - x_i'\beta_*)$$

$$(\beta_{L_2E} - \beta_*) = \left[\frac{1}{n}\sum_{i=1}^{n} H_\beta f_{a,b,c}(y_i - x_i'\beta)\right]^{-1}\left[-\frac{1}{n}\sum_{i=1}^{n} \bigtriangledown_\beta f_{a,b,c}(y_i - x_i'\beta_*)\right]$$

$$\sqrt{n}(\beta_{L_2E} - \beta_*) = \left[\frac{1}{n}\sum_{i=1}^{n} H_\beta f_{a,b,c}(y_i - x_i'\beta)\right]^{-1}\left[-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \bigtriangledown_\beta f_{a,b,c}(y_i - x_i'\beta_*)\right].$$

It can be shown that

$$\bigtriangledown_\beta f_{a,b,c}(y_i - x_i'\beta_*) = -f'_{a,b,c}(y_i - x_i'\beta_*)x_i,$$

where

$$f'_{a,b,c}(s) = \frac{\partial}{\partial s} f_{a,b,c}(s),$$

and

$$H_\beta f_{a,b,c}(y_i - x_i'\beta) = f''_{a,b,c}(y_i - x_i'\beta_*)(x_i x_i'),$$

where

$$f''_{a,b,c}(s) = \frac{\partial^2}{\partial s^2} f_{a,b,c}(s).$$

From these results, we can see that as $n \to \infty$,

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \bigtriangledown_\beta f_{a,b,c}(y_i - x_i'\beta_*) \to MN(\tilde{0}_{p+1}, \Sigma),$$

where

$$\begin{aligned}
\Sigma &= E[(\bigtriangledown_\beta f_{a,b,c}(Y - X'\beta_*))(\bigtriangledown_\beta f_{a,b,c}(Y - X'\beta_*))'] \\
&= E\left[\left(-f'_{a,b,c}(Y - X'\beta_*)X\right)\left(-f'_{a,b,c}(Y - X'\beta_*)X\right)'\right] \\
&= E\left[\left(f'_{a,b,c}(Y - X'\beta_*)\right)^2 (XX')\right] \\
&= E\left[\left(f'_{a,b,c}(\epsilon)\right)^2\right] E\left[(XX')\right].
\end{aligned}$$

Also,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} H_\beta f_{a,b,c}(y_i - x_i'\beta) \rightarrow E[H_\beta f_{a,b,c}(Y - X'\beta)] \\
&= E\left[f''_{a,b,c}(Y - X'\beta_*)(XX')\right] \\
&= E\left[f''_{a,b,c}(\epsilon)\right] E[(XX')] \\
&= V.
\end{aligned}$$

Putting all of these together, we see that

$$\sqrt{n}(\beta_{L_2E} - \beta_*) \rightarrow MN(\tilde{0}_{p+1}, V^{-1}\Sigma V^{-1}),$$

where the covariance matrix reduces to

$$\begin{aligned} V^{-1}\Sigma V^{-1} &= \left(E\left[f''_{a,b,c}(\epsilon)\right] E[(XX')]\right)^{-1} E\left[\left(f'_{a,b,c}(\epsilon)\right)^2\right] E\left[(XX')\right] \\ &\quad \times \left(E\left[f''_{a,b,c}(\epsilon)\right] E[(XX')]\right)^{-1} \\ &= \frac{E\left[\left(f'_{a,b,c}(\epsilon)\right)^2\right]}{E\left[f''_{a,b,c}(\epsilon)\right]^2} E[(XX')]^{-1} E[(XX')] E[(XX')]^{-1} \\ &= \frac{E\left[\left(f'_{a,b,c}(\epsilon)\right)^2\right]}{E\left[f''_{a,b,c}(\epsilon)\right]^2} E[(XX')]^{-1}. \end{aligned}$$

From this, given distributions for both X and the error, ϵ, we can determine the asymptotic covariance matrix of our regression coefficients. We can also estimate this result using our data to come up with an estimate of the covariance matrix of our regression coefficients, such as in Section 2.2.1.

3.3 Simulated Results

3.3.1 Quantile Estimates for Mixtures

To examine the effect of contamination on our L_2E quantile estimates, we examine three separate mixture densities. For each mixture, a range of weights, w, and parameters were examined in which 1000 points of data were simulated 1000 times, with $1000w$ points coming from the true density and $1000(1-w)$ points coming from the contamination density. A double exponential distribution with $a = b = 1$ was used for the L_2E criteria, estimating the median of each distribution. The average minimum over the simulations for each combination was then used to make the contour plots in Figure 3.4. In each contour plot, the areas where there are two theoretic minima are shaded blue, while the areas where there is a single theoretic minimum are shaded

red.

The first mixture, shown in Figure 3.4(a), is a mixture of a Unif(0,1) distribution, considered the true density, and a Unif(d, $d+1$) distribution, where $d > 1$ is the left endpoint of the contamination density. That is, the mixture density can be represented by

$$w\text{Unif}(0,1) + (1-w)\text{Unif}(d, d+1).$$

The region to the right of the 0.51 contour line are combinations of the weight and the contamination left endpoint that have simulated means of 0.50 ± 0.01. We can see that as w and d increase, the bias added to the estimated values of θ_{L_2E} by the contamination density decreases. In particular, once there is a large enough difference between the two distributions in the mixture, the estimate within the uncontaminated density is not very affected by the contamination density, regardless of the weight.

The second mixture, shown in Figure 3.4(b), is a mixture of a Unif(0,1) distribution, again considered the true density, and a Unif(1,$s+1$) distribution, where $s > 0$ is the width of the contamination density. Thus, the mixture density can be represented by

$$w\text{Unif}(0,1) + (1-w)\text{Unif}(1, 1+s).$$

Though not as drastic as the previous mixture, we see that increasing the parameters w and s decreases the bias on the estimated values of θ_{L_2E}. We can also see from this mixture is that if there is no separation between the two densities in the mixture, we will see an effect on the estimate within the uncontaminated region, adding bias towards the contamination.

The third mixture, shown in Figure 3.4(c), is a mixture of a N(0,1) distribution, considered the true density, and a N(μ,1) distribution, considered the contamination

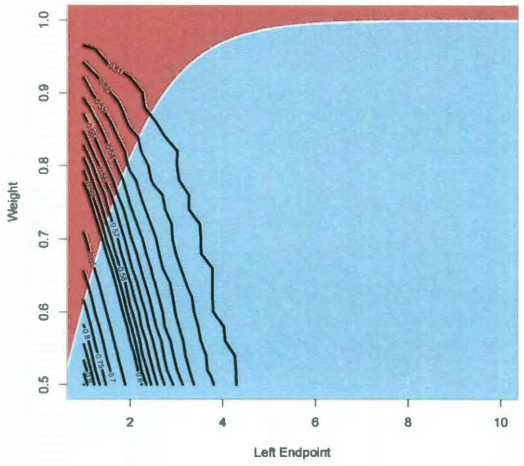
density. This mixture can be represented by

$$wN(0,1) + (1-w)N(\mu,1).$$

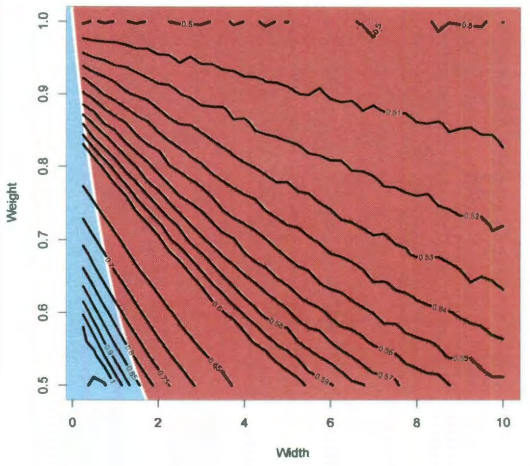
The region to the right of the 0.01 contour line are combinations of the weight and the contamination mean that have simulated means of 0 ± 0.01. Once again, by increasing w and μ, the bias caused by the contamination density on the estimated values of θ_{L_2E} decreases. As before in the first mixture, when there is a large enough separation between the two densities, the estimate within the uncontaminated density is not very affected by the contamination density, regardless of the weight.

3.3.2 Standard Deviation

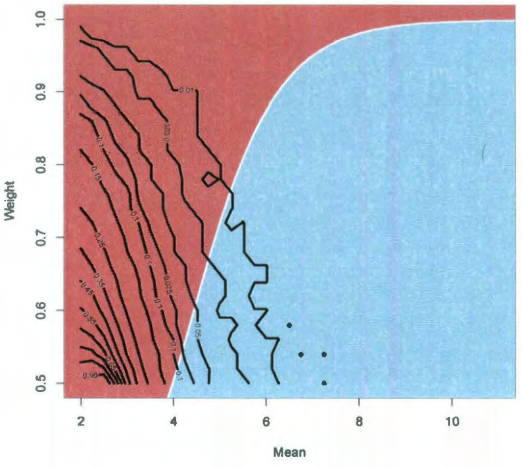
To verify the theoretic standard deviations found in Section 3.2, a range of combinations of a and b we simulate 10,000 samples of Uniform(0,1) data of size 1000. We keep track of the estimated values of θ_{L_2E} for each sample. The standard deviations of these estimated values of θ_{L_2E} for each pair of a and b can be found in Figure 3.5(a). We repeat this process using Normal(0,1) data of size 1000, the results of which can be found in Figure 3.5(b). As we can see, these simulated results closely match both the theoretic values and the trends of those values found in Figure 3.3. This lends credence to our formulas for asymptotic behavior of θ_{L_2E}. One trend of note that is featured in both plots is the monotonicity of the standard deviation across the median, that is, when $a = b$, as $a + b$, increases. This leads us to believe that smaller values of r cause our estimate to have a smaller standard deviation. However, this is not the only consideration to be taken into account when selecting a value of r.



(a) $w\text{Unif}(0,1) + (1-w)\text{Unif}(\text{LE}, \text{LE}+1)$

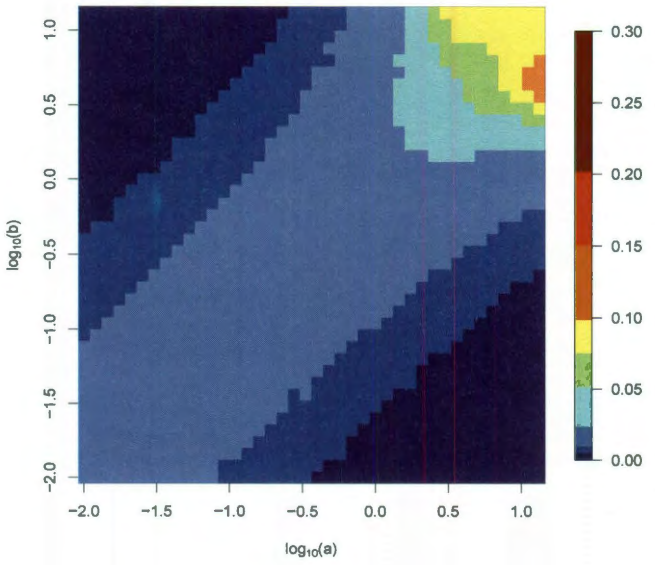


(b) $w\text{Unif}(0,1) + (1-w)\text{Unif}(1, 1+\text{Width})$

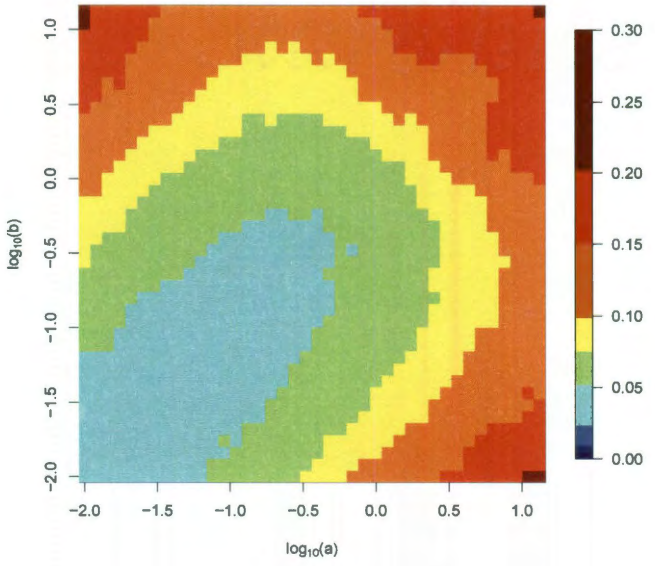


(c) $wN(0,1) + (1-w)N(\mu, 1)$

Figure 3.4 : Theoretic quantiles estimated from various mixture models. The red regions indicate L_2E criteria with one theoretic minimum while blue regions indicate criteria with two theoretic minima.



(a) Uniform



(b) Normal

Figure 3.5 : Simulated standard deviations for L_2E quantile estimates given various values of a and b and a sample size of 1000.

3.4 Choosing a Value of r

As mentioned in Section 3.1, to find unique solutions to our L_2E minimization in Equation 3.1, it is necessary to set $a+b$ equal to some constant r. Ideally, this value of r minimizes the variance for all values of τ, giving us as accurate of an estimator as possible. For our purposes, we will examine the case where $g(x) = N(0,1)$. As shown by the curves in Figure 3.6, variance goes down as r goes to 0. This property is illustrated in Figure 3.7, where the standard deviations at each quantile for traces of $r = .01, 1, 2,$ and 10 are shown.

However, we still need a value of r that gives us good numeric results when we are using our L_2E quantile estimator. In Figure 3.8, we again see the theoretic quantile contours for $N(0,1)$ data. Curves representing $a+b=r$ are plotted on the contours, where $r = .01, 1, 2,$ and 10. As we can see, if we have too small of a value of r, it can be difficult for extreme quantiles to be estimated, as either the value of a or b has to be so small that it can affect the optimization. Because we want the smallest value of r possible that can estimate the extreme quantiles without having a or b be too close to 0, we propose that an r somewhere around 2 is a reasonable choice.

3.5 Discussion

As we can see, given the parameters a and b for our L_2E quantile estimation criteria and an assumed distribution $g(x)$ for the data, we are able to determine which theoretic quantile values are being estimated. Likewise, we are able to choose a and b to achieve a specific quantile level, τ, such that the theoretic quantile estimate θ_{L_2E}, is equal to $G^{-1}(\tau)$. We can also determine the asymptotic behavior of our estimate, allowing us to determine its distribution. The asymptotic behavior, along with the step

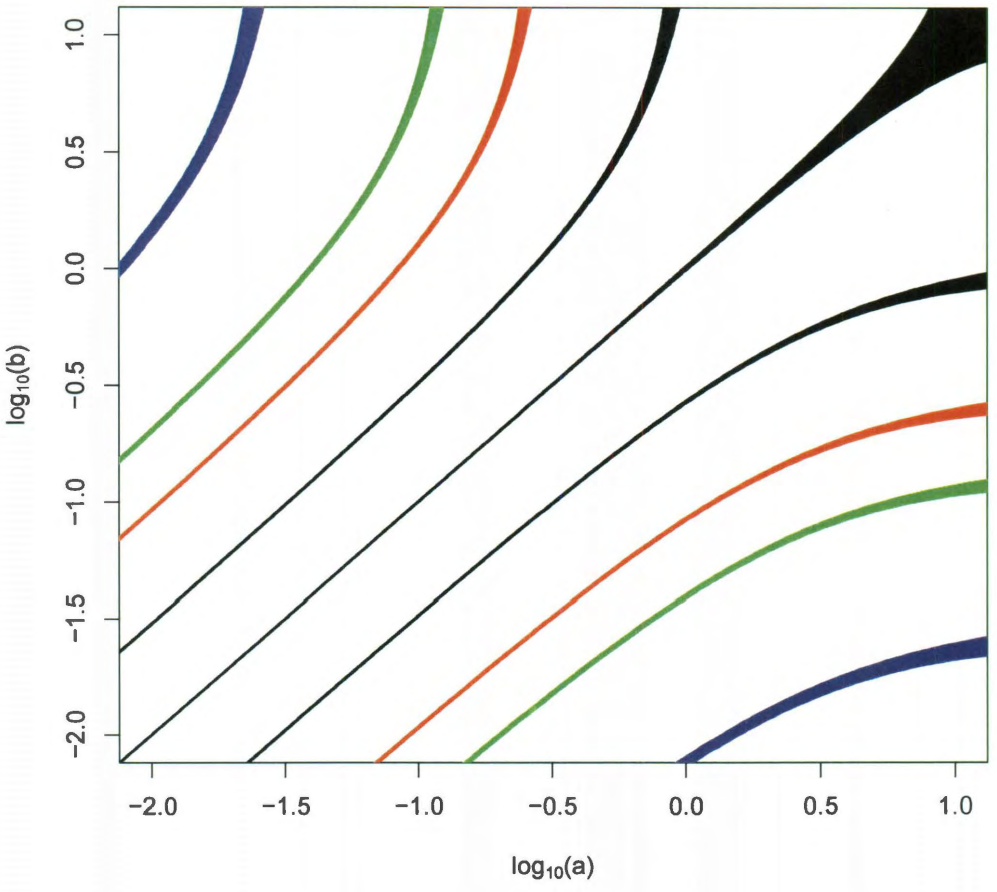
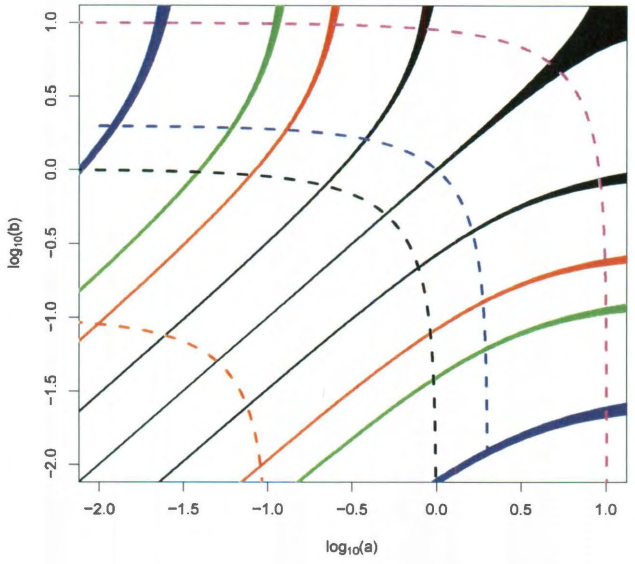
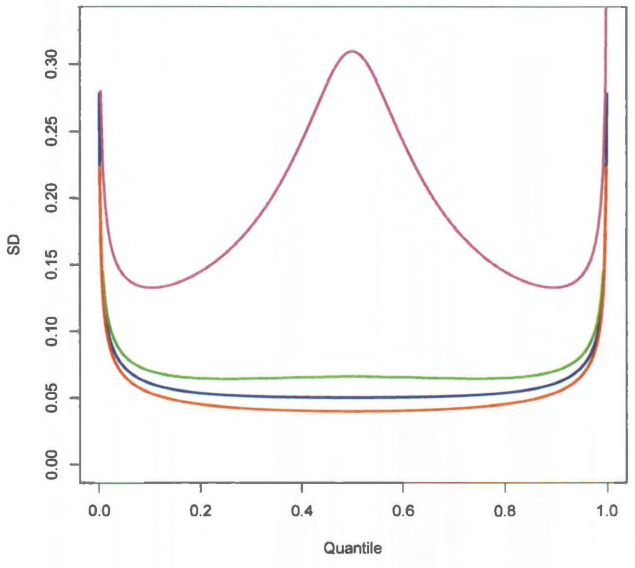


Figure 3.6 : Standard deviations for each quantile curve. At each point on the curve, the width of the curve is proportional to the standard deviation of θ_{L_2E} for those values of a and b. From the bottom, the .01, .05, .1, .25, .5, .75, .9, .95, and .99 quantile levels are shown.



(a)



(b)

Figure 3.7 : Taking $a + b = r$, the values over the traces seen in (a) can be found in (b). From bottom to top in (b), $r = .01$, 1, 2, and 10.

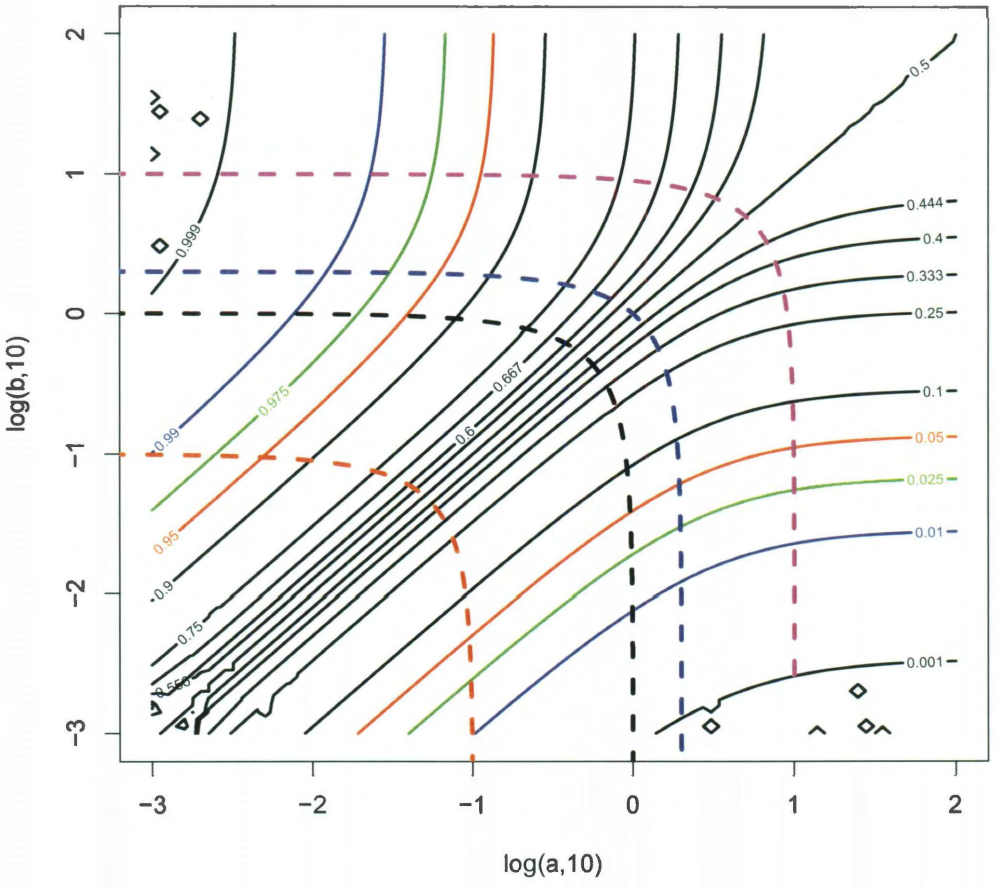


Figure 3.8 : Standard deviations for each quantile curve. At each point on the curve, the width of the curve is proportional to the standard deviation of θ_{L_2E} for those values of a and b. From the bottom, the .01, .05, .1, .25, .5, .75, .9, .95, and .99 quantile levels are shown.

size of optimization methods, allow us to determine a good value of r, to constrain the value of $a + b$ and give us unique choices of the parameters to achieve specific theoretic quantiles.

Chapter 4

Non-linear and Semiparametric Robust Quantile Regression

Up to now, our focus has been on linear L_2E quantile regression. However, as with other regression criteria, other forms of regression can be extended from our L_2E criterion, both non-linear and semiparametric. A simple non-linear example would be quadratic regression. Given a bivariate sample $(x_1, y_1), \ldots, (x_n, y_n)$, we can find the coefficients of a quadratic L_2E quantile regression line by solving the minimization

$$\arg\min_{\beta} \int f_{a,b,c}(x)^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f_{a,b,c}\left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2\right).$$

There are many potential extensions to L_2E quantile regression, including penalized methods and autoregressive processes, among others. In this chapter, we discuss two possible semiparametric extensions to L_2E quantile regression: polynomial splines and local polynomial regression. Examples of each extension are also shown.

4.1 Quantile Regression with Polynomial Splines

A possible extension to L_2E quantile regression is polynomial splines. These splines are used for smoothing and are appropriate when the functional relationship between the predicting and response variables is unknown. One method of implementing polynomial splines is to use a truncated power basis of degree p with K selected

knots, or points where the polynomial function is thought to change.That is, given a bivariate sample $(x_1, y_1), \ldots, (x_n, y_n)$ and knots placed at $(\kappa_1, \ldots, \kappa_K)$, we can find a spline fit of degree p by solving the minimization

$$\arg\min_{\beta} \int f_{a,b,c}(x)^2 \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} f_{a,b,c} \left(y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_p x_i^p - \sum_{k=1}^{K} \beta_{p+k} (x_i - \kappa_k)_+^p \right),$$

where the so-called hinge function is defined as

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

So, for example, cubic splines, that is, degree 3, can be found by the minimization

$$\arg\min_{\beta} \int f_{a,b,c}(x)^2 \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} f_{a,b,c} \left(y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_3 x_i^3 - \sum_{k=1}^{K} \beta_{3+k} (x_i - \kappa_k)_+^3 \right).$$

Like least squares polynomial splines, this particular method can likely be improved through the use of penalization and other bases than the truncated power basis, such as the B-spline basis. However, for our current purposes, we use the method as presented above.

4.1.1 Example

We simulate 900 points of data, $(x_1, x_2, \ldots, x_{900})$, from a normal distribution with $\mu = 11$ and $\sigma = 2.5$. From these points, $(y_1, y_2, \ldots, y_{900})$ are derived from the equation

$$y_i = x_i sin(x_i) + \epsilon_i,$$

where ϵ_i are simulated from a normal distribution with $\mu = 0$ and $\sigma = 2$. The pairs, $(x_1, y_1), (x_2, y_2), \ldots, (x_{900}, y_{900})$, are then considered to be the uncontaminated data. A cluster of 100 points of simulated multivariate normal data located above the uncontaminated data are then added to the full data set and are considered to be the contaminated data. That is, we have 1000 points of data with 10% contamination.

Following the method described above, linear splines are fitted to the data in Figure 4.1 and cubic splines are fitted to the data in Figure 4.2. For the linear splines, knots were selected to be at 5, 8, 11, and 14 while for the cubic splines, knots were selected to be at 8, 11, and 14. The .05, .10, .25, .50, .75, .90, and .95 quantile estimates are shown for each set of splines. As we can see in both cases, the quantile estimates are mostly unaffected by the contaminated data, showing good robustness. It is worth noting that these splines do not guarantee monotonicity among the curves, particularly in sparse regions, as evidenced by the cubic splines. Adding more or changing the knots can fix this issue, if need be. The behavior of these cubic splines over many simulations can be found in Section 5.2.

4.2 Local Polynomial Quantile Regression

Another extension to L_2E quantile regression is local polynomial regression. The idea behind this form of regression is that given a point, x_0, for the conditional estimate to be calculated, points close to x_0 are given more weight than points far away. In standard polynomial with degree p quantile regression, by rewriting Equation 2.2 as

$$\arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(\int f_{a,b,c}(x)^2 \mathrm{d}x - 2 f_{a,b,c}(y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_p x_i^p) \right),$$

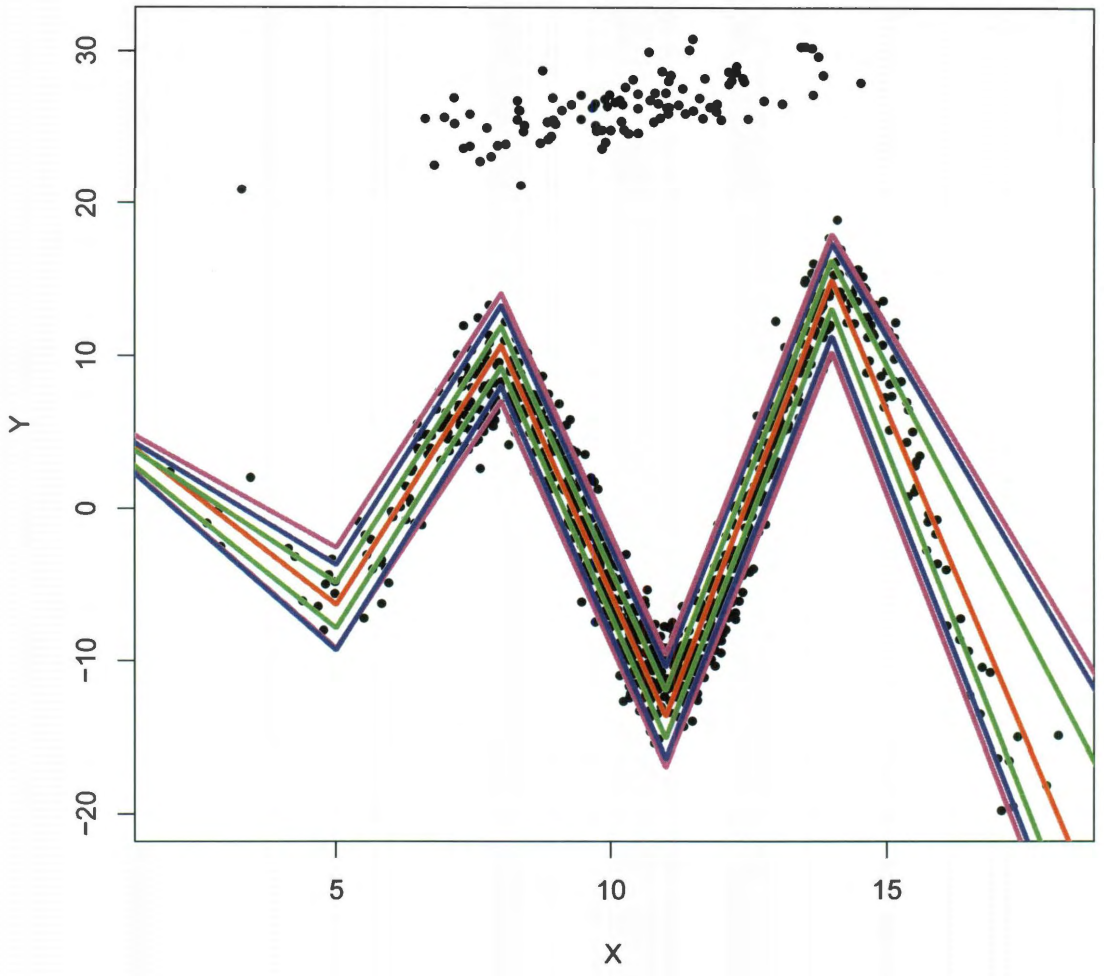


Figure 4.1 : L_2E quantile regression with linear splines. Knots are placed at 5, 8, 11 and 14. The .05, .10, .25, .50, .75, .90, and .95 quantile levels are shown.

we see that the amount of information each point gives to an estimate is $\int f_{a,b,c}(x)^2 \mathrm{d}x - 2f_{a,b,c}(y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_p x_i^p)$. From this, we can then perform local polynomial quantile regression to find the conditional quantile estimate of Y given $X = x_0$ by

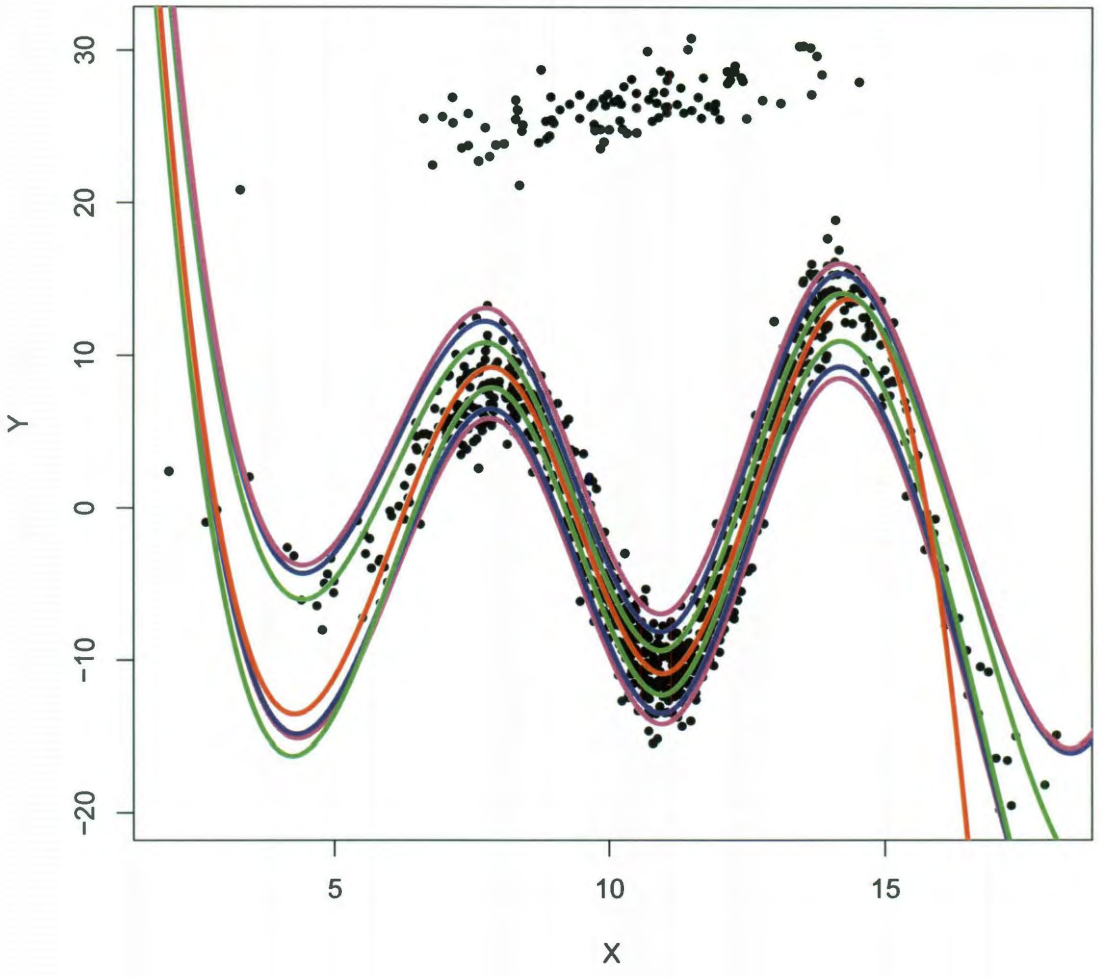


Figure 4.2 : L_2E quantile regression with cubic splines. Knots are placed at 8, 11, and 14. The .05, .10, .25, .50, .75, .90, and .95 quantile levels are shown.

minimizing

$$\arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (\int f_{a,b,c}(x)^2 \mathrm{d}x - 2 f_{a,b,c}(y_i - \beta_0 - \beta_1(x_i - x_0) - \cdots - \beta_p(x_i - x_0)^p)) K\left(\frac{x_i - x_0}{h}\right),$$

where K(x) is a specified kernel density. The conditional quantile estimate is then $\hat{\beta}_0$. The kernel densities are, in general, positive symmetric functions with decreasing values as $|x| \to \infty$, such as a standard normal distribution. The bandwidth of the kernel, h, is a positive value and can be selected through trial and error. The choice of p also must be selected, with values of $p = 1$ or 2 usually acceptable. This method seems particularly useful for data with no clear functional form. However, because each point must be calculated through a separate minimization, finding each estimated quantile curve takes considerably longer than polynomial splines.

4.2.1 Example

Using the simulated data from before in Section 4.1.1, an example of L_2E local linear quantile regression can be found in Figure 4.3. For this example, a standard normal kernel is used with a bandwidth of $h = 1/3$. As before, the .05, .10, .25, .50, .75, .90, and .95 conditional quantile estimates are shown. We see that these curves are similar to the cubic spline conditional quantile curves in Figure 4.2, with the added benefit of monotonicity of the conditional quantile curves. Once again, we also see that this method displays robustness, as it is seemingly unaffected by the contaminated data.

4.3 Discussion

As we have seen, we can extend our L_2E quantile regression criteria to perform other forms of regression than linear regression. In particular, we are able to perform quantile regression using polynomial splines and local polynomial regression to obtain conditional quantile estimates. Future research is needed to determine the theoretic and asymptotic behavior of these extensions, as well as adapting our criteria for other extensions.

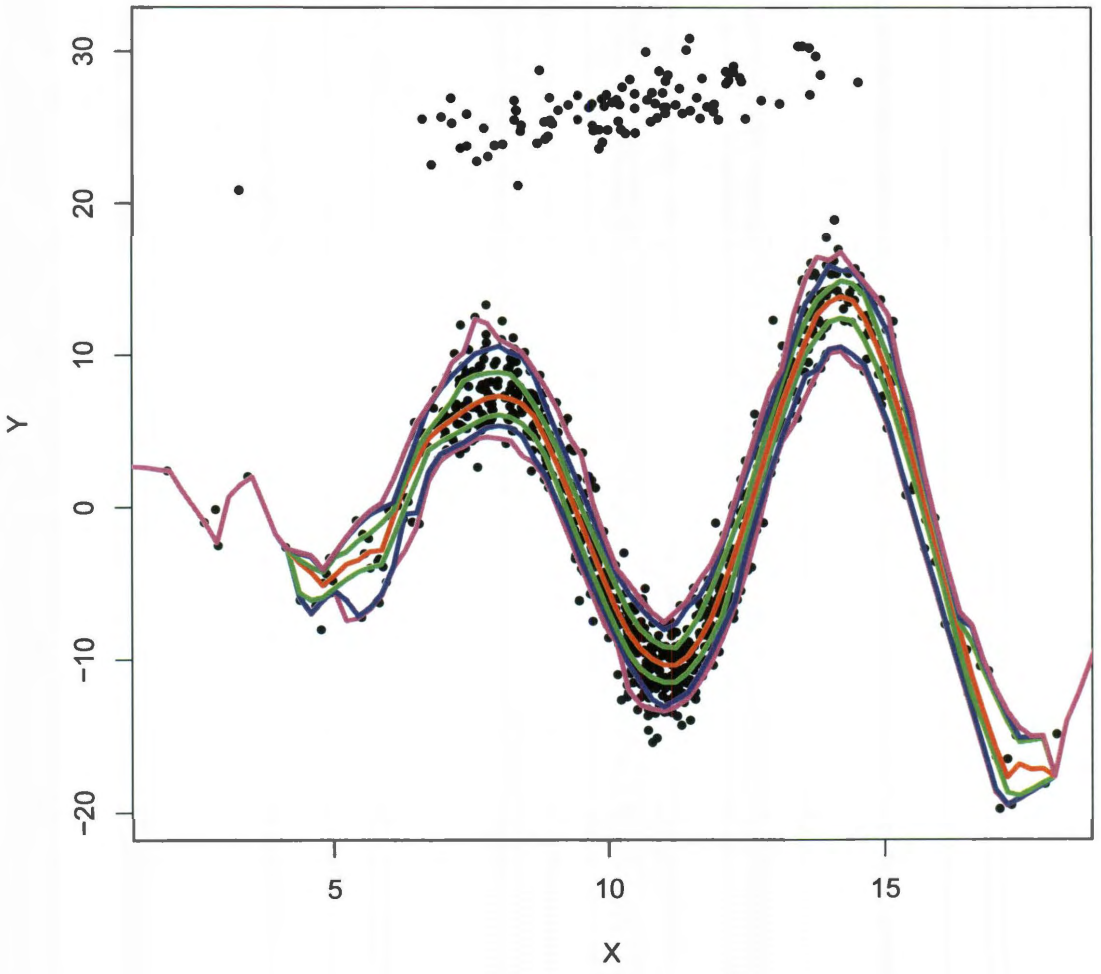


Figure 4.3 : L_2E local linear quantile regression. The .05, .10, .25, .50, .75, .90, and .95 quantile levels are shown.

Chapter 5

Analysis of Simulated Data

Before using L_2E quantile regression on real data, it is worthwhile to examine the effects of our methods on much more controlled environments. In particular, we need to examine how L_2E quantile regression behaves when using simulated data, in terms of both robustness and accuracy. We also need to check the accuracy and reliability of our model selection criteria. By doing these things, we can then examine real data with confidence in our methods.

5.1 Data with Normal Residuals and Contamination

To examine the added robustness provided by L_2E quantile regression, we compare the quantile regression lines estimated by our L_2E method described in Chapter 2 with those estimated by KB's method. To do this, 900 points of bivariate normal data are simulated. These pairs of data are considered to be the uncontaminated data. This data is such that the residuals around the least squares regression line are distributed $N(0, \sigma^2)$, where σ is unknown. Then, 100 points of bivariate normal data are simulated, such that they are centered above the uncontaminated data cloud. These pairs of data are then added to the uncontaminated data and are considered the contaminated data. Thus, we have 1000 points of data with 10% contamination. Then, we estimate the linear quantile regression lines using our L_2E criteria on the full data set, KB's criteria on the full data set, and then KB's criteria on just the un-

contaminated data set. An example of these regression lines, in particular estimating the .01, .05, .1, .25, .5, .75, .9, .95, and .99 quantiles, can be found in Figure 5.1.

As we can see, the quantile regression lines estimated using L_2E match up fairly well with the quantile regression lines estimated using KB's method on just the uncontaminated data. As seen before, the upper quantile regression lines, in particular, the .9, .95 and .99 quantiles, jump up to the contamination data when using KB's method on the full data set, while they stay within the uncontaminated data when using L_2E.

τ	L_2E	$\hat{\sigma}_{L_2E}$	KB (F)	$\hat{\sigma}_{Kf}$	KB (UC)	$\hat{\sigma}_{Kuc}$
.01	.011	.002	.011	.001	.010	.001
.05	.054	.005	.056	.001	.050	.001
.10	.105	.007	.112	.001	.100	.001
.25	.253	.012	.278	.001	.251	.001
.50	.500	.014	.556	.001	.501	.001
.75	.750	.012	.834	.001	.751	.001
.90	.900	.006	.986	.002	.901	.001
.95	.950	.004	.999	.001	.951	.001
.99	.991	.002	1.00	.000	.991	.001

Table 5.1 : Summary of Quantile Results For $N(0, \sigma^2)$ Residuals From 1000 Simulations

One method of determining what quantile level a quantile regression line is estimating is to determine the proportion of residuals about the regression line that are negative. That is, we want to know what proportion of y values are smaller than the estimated conditional quantiles. Another way to look at this is that if we want a level τ quantile regression line, then the proportion of negative residuals about the regression line should be τ. We use this idea to test the robustness and accuracy of the quantile regression methods. To do this, we simulate the above data 1000 times and keep track of the proportion of negative residuals for each estimated quantile line

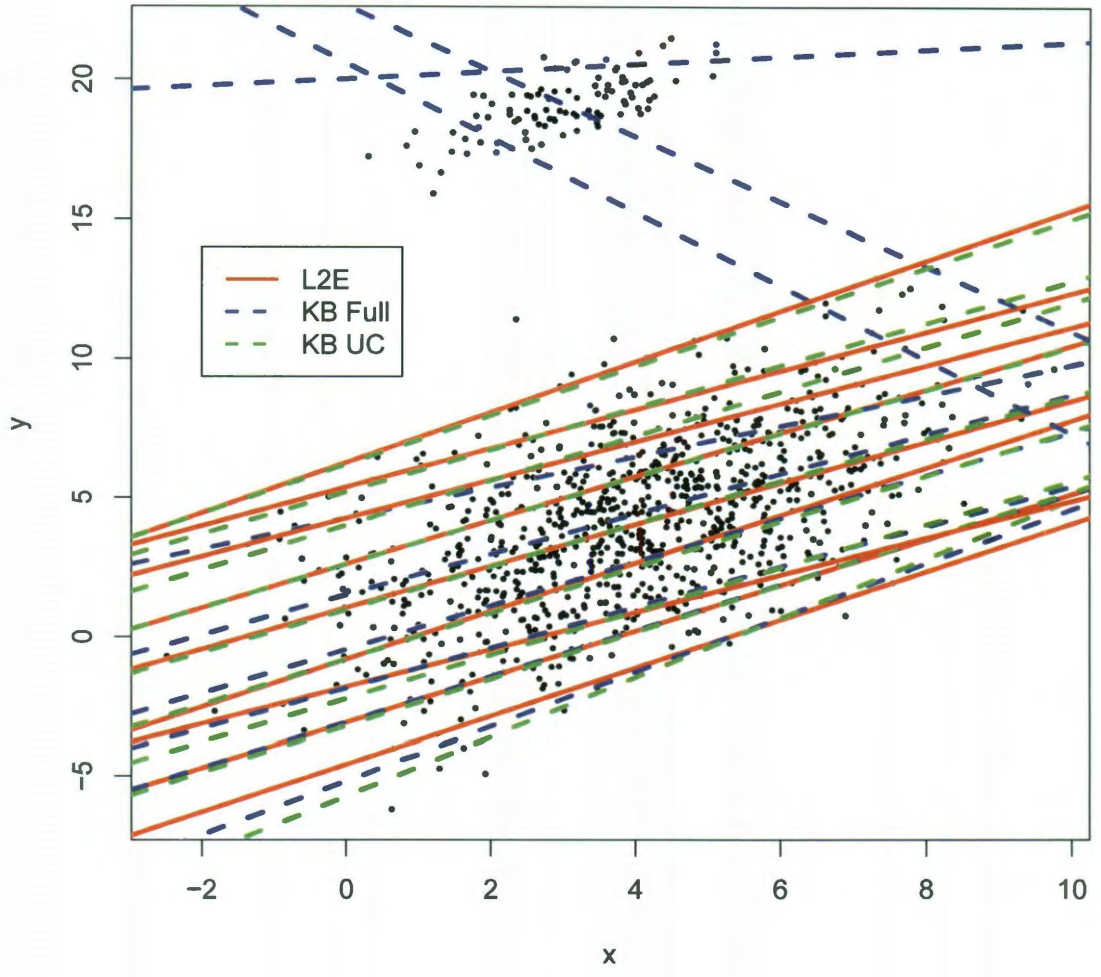


Figure 5.1 : Comparison of L_2E quantile regression, shown in red, KB's quantile regression on the full data set, shown in blue, and KB's quantile regression on the uncontaminated data on 900 points of bivariate normal data with 100 points of contamination added above. Least squares residuals are assumed to be $N(0, \sigma^2)$, where σ is unknown.

for each method. In Table 5.1, the mean and standard deviations of each of these proportions is shown. We can see that the L_2E method not only matches closely to the results found by using KB's method on the uncontaminated data, it seems relatively unaffected by the contaminated data, unlike the results from using KB's method on the full data set. Thus, we see the affects that using L_2E for quantile regression has on the robustness of these conditional quantile estimates.

5.2 Sinusoidal Data with Contamination

Similarly, to see the effect of the added robustness on conditional quantile estimation using cubic splines, we simulate 1000 pairs of data using the same method described in Section 4.1.1. That is, 900 points of data, $(x_1, x_2, \ldots, x_{900})$, are simulated from a normal distribution with $\mu = 11$ and $\sigma = 2.5$. From these points, $(y_1, y_2, \ldots, y_{900})$ are derived from the equation

$$y_i = x_i sin(x_i) + \epsilon_i,$$

where ϵ_i are simulated from a normal distribution with $\mu = 0$ and $\sigma = 2$. The pairs, $(x_1, y_1), (x_2, y_2), \ldots, (x_{900}, y_{900})$, are then considered to be the uncontaminated data. 100 points of simulated multivariate normal data placed above the uncontaminated data are then added to the full data set and are considered to be the contaminated data. That is, we have 1000 points of data with 90% uncontaminated sinusoidal data and 10% multivariate normal contamination. Once again, we compare the results of using cubics splines with our L_2E quantile regression criteria on the full data set to using cubic splines with KB's asymmetric absolute loss quantile regression criteria on both the full data set and also just the uncontaminated data. An example of this comparison of the two methods on the uncontaminated data can be seen in Figure

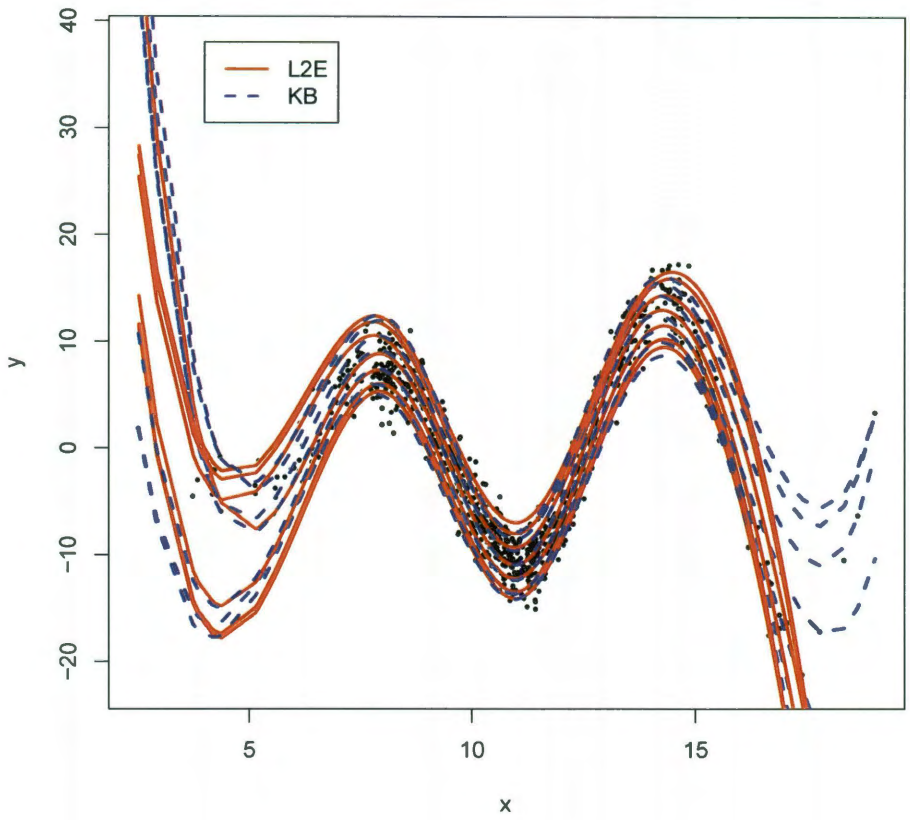


Figure 5.2 : Comparison of cubic splines with L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the uncontaminated sinusoidal data. Residuals about the L_2E median cubic splines are assumed to be $N(0, \sigma^2)$, where σ is unknown.

5.2, where the .05, .1, .25, .5, .75, .9, and .95 conditional quantile curves estimated by each method are shown. Knots were placed at 8, 11, and 14. Likewise, a comparison of the two methods on the full data set can be seen in Figure 5.3.

Once again, the estimated quantiles are found by looking at the proportion of the residuals that are negative. That is, the true y values that have values smaller

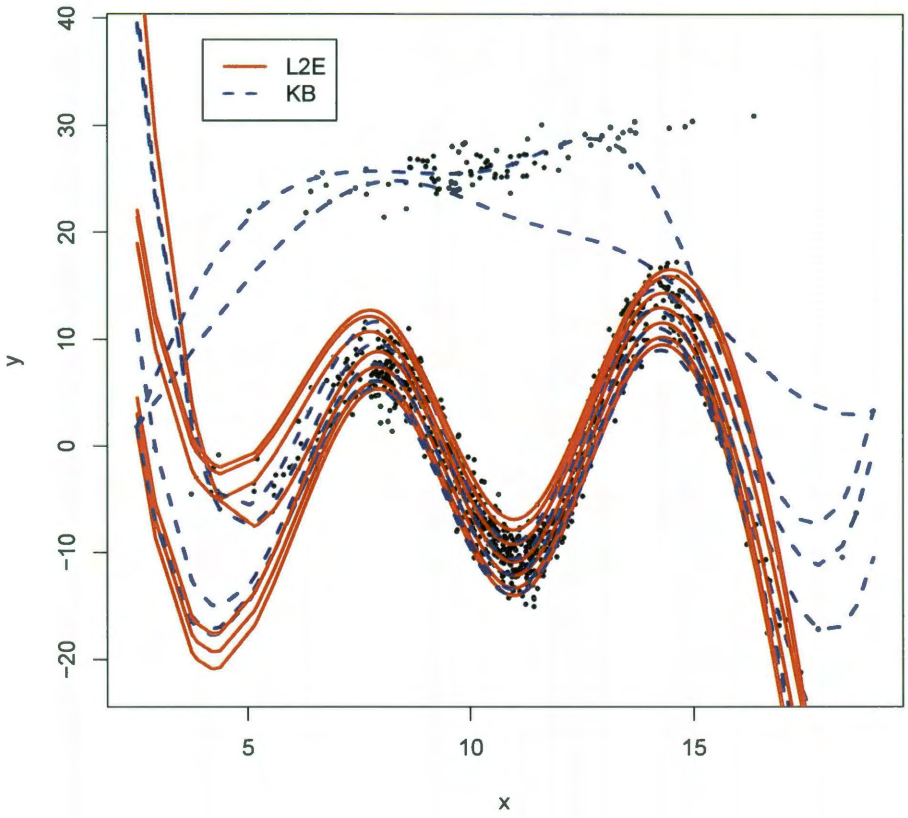


Figure 5.3 : Comparison of cubic splines with L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the sinusoidal data set with contamination. Residuals from the uncontaminated data about the L_2E median cubic splines are assumed to be $N(0, \sigma^2)$, where σ is unknown.

than the estimated conditional quantile values. To examine the robustness of each method, we only look at the residuals from the uncontaminated data. In Table 5.2, the average quantile estimated by each of these methods over 1000 simulations is shown for a range of desired quantiles, along with the standard deviation of these estimated quantiles. As we can see, using L_2E quantile regression is far less affected

by the outlying data, particularly in the upper quantiles. This leads us to believe that by using L_2E, there is a clear level of added robustness when using cubic splines.

τ	L_2E	$\hat{\sigma}_{L_2E}$	KB (F)	$\hat{\sigma}_{Kf}$	KB (UC)	$\hat{\sigma}_{Kuc}$
.05	.061	.008	.056	.003	.050	.003
.10	.111	.010	.111	.003	.100	.003
.25	.258	.011	.278	.003	.250	.003
.50	.499	.009	.555	.003	.500	.003
.75	.747	.011	.833	.003	.750	.003
.90	.893	.010	.982	.003	.900	.003
.95	.942	.009	.995	.002	.950	.003

Table 5.2 : Summary of Quantile Results For Cubic Spline Residuals From 1000 Simulations

5.3 Model Selection

In order to test the model selection criterion described in Section 2.2.2, that is, the AIC criterion, 100 points of data were simulated for each of the following six $N(0, \sigma^2)$ distributions:

$$X_1 \sim N(0, 1^2) \qquad X_2 \sim N(1, .5^2) \qquad X_3 \sim N(6, 1^2)$$

$$X_4 \sim N(0, 2^2) \qquad X_5 \sim N(1, .3^2) \qquad X_6 \sim N(0, 2^2)$$

From these distributions, a response variable, Y, is created by

$$Y = 3X_1 - 2X_3 + 4X_6 + \epsilon,$$

where $\epsilon \sim N(0, .5^2)$. From these 100 sets of points, a best-subset analysis is run using all linear combinations of the six predictor variables. For each $p \in 1, 2, \ldots, 6$, the

combination of p predictor variables with the lowest AIC value is stored. From these combinations, the combination with the lowest AIC value is selected as being the best reduced model.

Using the same distributions for each variable, 1000 data sets are simulated and the the best reduced model is selected for each. The simulations were then repeated for sample sizes of 1000 and 2500. From these simulations, the proportion of times that a model with p predictor variables was selected can be found in the table below:

p	1	2	3	4	5	6
$n = 100$	0.00	0.00	0.083	0.328	0.421	0.168
$n = 1000$	0.00	0.00	0.066	0.312	0.417	0.205
$n = 2500$	0.00	0.00	0.082	0.296	0.427	0.195

As we can see, AIC tends to be conservative in the model selection, choosing a model with either four or five factors about 75% of the time for all three sample sizes. Models with two or fewer variables were not selected in any of the simulations of all three sample sizes. To see which variables were selected, the proportion of times that each predictor variable were included in the best model can be found in the table below:

Variable	X_1	X_2	X_3	X_4	X_5	X_6
$n = 100$	1.000	0.566	0.999	0.551	0.558	1.000
$n = 1000$	1.000	0.600	1.000	0.574	0.587	1.000
$n = 2500$	1.000	0.579	1.000	0.581	0.575	1.000

We see that the three true predictor variables, X_1, X_3, and X_6 were included in in the best model from almost every simulation for all three sample sizes. The other three variables were each included about 55-60% of the time. Another point of interest is which variables were included in the best model with three predictor variables. The

proportion of time each predictor variable was included in these models can be seen below:

Variable	X_1	X_2	X_3	X_4	X_5	X_6
$n = 100$	1.00	0.00	1.00	0.00	0.00	1.00
$n = 1000$	1.00	0.00	1.00	0.00	0.00	1.00
$n = 2500$	1.00	0.00	1.00	0.00	0.00	1.00

Once again, the true predictor variables were part of the best best model with three predictor variables in every simulation for all three sample sizes. This gives us confidence in our criteria. In fact, in such cases where models with more than three predictor variables was selected, the difference in AIC values is quite small. For example, in the table below, where a Y indicates that the variable was included in the p parameter model and an N indicates that it was not, a five variable model was selected. A sample size of 100 was used.

p	X_1	X_2	X_3	X_4	X_5	X_6	AIC
1	N	N	N	N	N	Y	1555.880
2	Y	N	N	N	N	Y	-59.849
3	Y	N	Y	N	N	Y	-207.673
4	Y	N	Y	Y	N	Y	-208.072
5	Y	N	Y	Y	Y	Y	-208.315
6	Y	Y	Y	Y	Y	Y	-208.203

It is clear that the difference in the best five parameter model is only marginally better than the best three parameter model. In fact, there isn't much difference in the AIC values between the best models of $p = 3$, 4, 5, or 6. However, there is a noticeable difference in the AIC values of the one and two parameter models. This leads us to believe that a best subsets analysis is useful in selecting a model, as it is

possible to find the model with the fewest predictor variables that has an AIC value that is either the lowest or only marginally higher than the lowest.

5.4 Discussion

These simulated results show that both our linear and semiparametric L_2E quantile regression behave as expected, in that they are estimating the desired quantile levels of the uncontaminated data. Thus, in both cases we see the added level of robustness from our L_2E criteria over using KB's criteria. We also see that our model selection using a robust AIC is useful in model reduction. Because these simulated results behave as expected on simulated data, we feel confident in using these methods on real data.

Chapter 6

Analysis of Real Data

In this chapter, we examine the behavior of L_2E quantile regression on real data. We begin by analyzing data that was already analyzed by Koenker using his and Bassett's method to compare the results. We then analyze a data set with a non-linear trend. Finally, we analyze a data set that shows an application of quantile regression and allows us to test our AIC model selection criteria on real data.

6.1 Birth Weight Data

In order to compare the effects of L_2E quantile regression to KB's quantile regression on a real data set, we replicate the study performed by Koenker and Hallock (2001) on the impact of various factors on the birthweight of infants born in the United States. Their study, in turn, was based on an analysis of the same topic by Abrevaya (2001). These studies sought to find which factors can be attributed to lower infant birthweights, thus allowing for measures to be taken to reduce the number of low-birthweight incidents.

The data comes from the June 1997 Detailed Natality Data collected by the National Center for Health Sciences. Abrevaya, and thus Koenker and Hallock, restricted the sample to live, singleton births where the mother was between the ages of 18 and 45, residing in the United States, and either black or white. Incomplete observations were ignored. This lead to a sample of size n =198,377, with 15 predictor variables

and one response variable. The response variable, birthweight, is measured in kilograms, unlike the study by Koenker and Hallock were it was measured in grams. A cursory look at the data leads us to believe that there are not are large number of outliers in the data.

The predictor variables include the mother's age (in years) and the mother's weight gain (in pounds). Both of these variables are included in the model as quadratic effects. Indicator variables are included, where a 1 indicates yes, designating if the child was a boy, if the mother was black, if the mother was married, and if the mother was a smoker. The number of cigarettes the mother smoked each day is also included. Education is split into indicator variables of which high school graduate, some college, and college graduate are included in the model. The omitted category is less than high school graduate. Likewise, the time of the first prenatal visit is split into indicator variables. No prenatal visits, first visit during the second trimester, and first visit during the third trimester are included in the model, while first visit during the first trimester is omitted.

For this study, regression lines were estimated using L_2E quantile regression for each of the 21 values of τ in the range $\{.01, .05, .10, \ldots, .90, .95, .99\}$. The coefficient estimates from each regression line were stored and then plotted in red against the values of τ, as seen in Figure 6.1. The shaded region in these plots represents the estimated 95% confidence region for each estimated coefficient. This process was then repeated using KB's quantile regression. Those coefficient estimates can be seen in blue. A least squares regression line was also estimated, with its coefficient estimates and corresponding 95% confidence intervals represented by the dashed black lines.

As we can see, the coefficient estimates from our L_2E method are very similar to the coefficient estimates from KB's method. In fact, most of the coefficient estimates

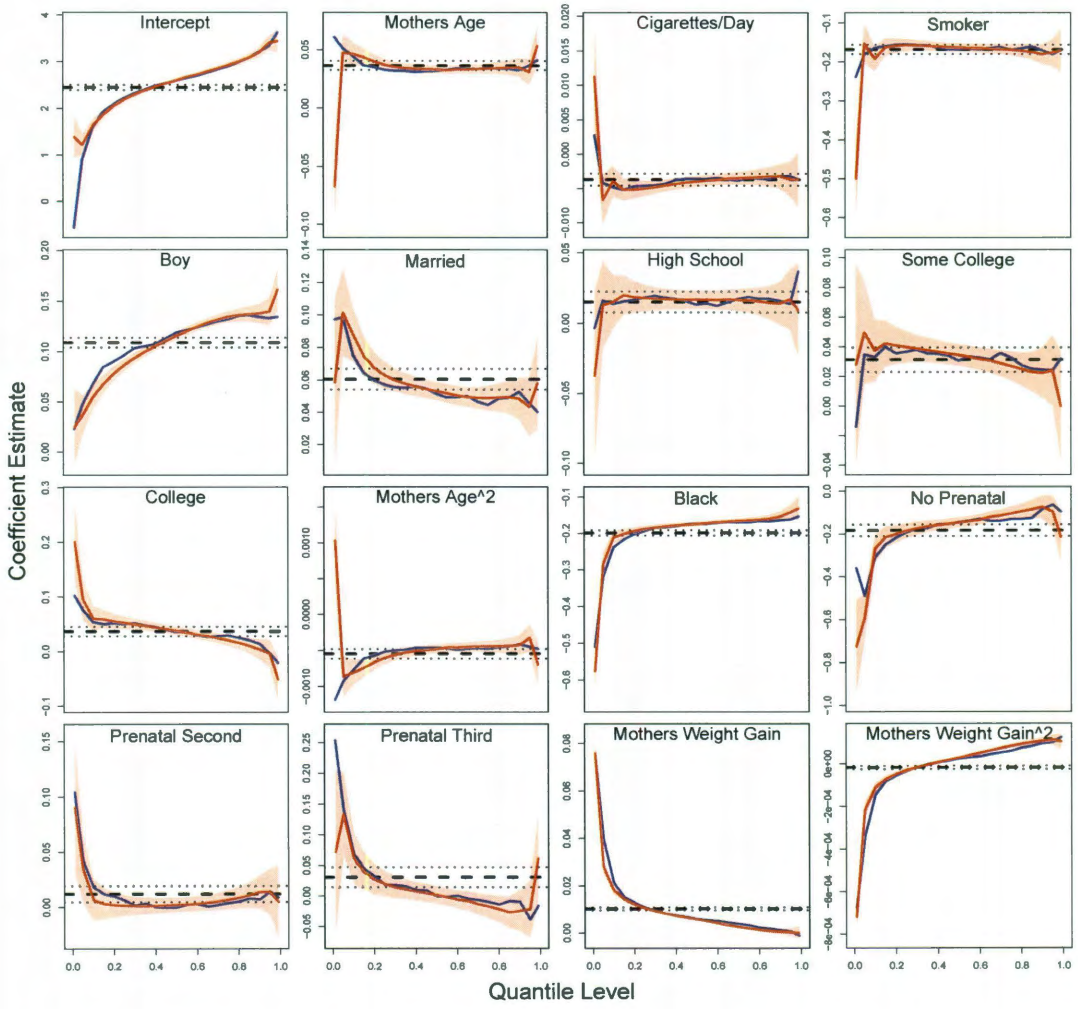


Figure 6.1 : Comparison of coefficient estimates from L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the birth weight data set. The shaded red regions represent the 95% confidence intervals for the L_2E quantile regression coefficient estimates. The least squares coefficient estimates, along with their 95% confidence interval, are shown in black.

from KB's quantile regression fall within the confidence interval around the L_2E coefficient estimates. The estimates that differ greatly tend to come from the extreme

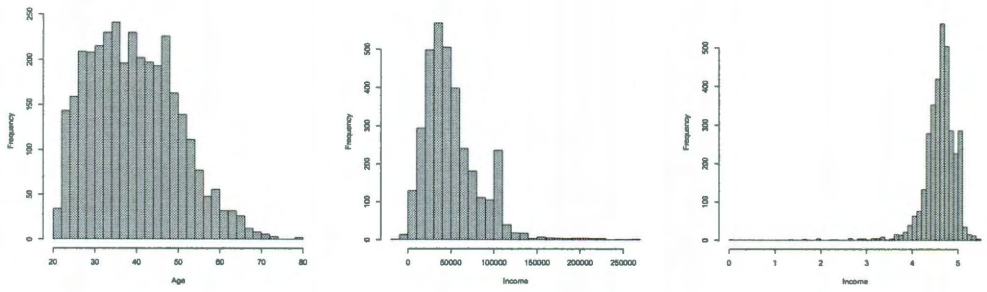
quantile levels, in particular where $\tau = .01$ or .99. These differences likely come about due to some instability in the numerical optimization. The coefficient estimates being nearly equivalent gives us confidence in our assumption that there are not a significant number of outliers in the dataset, as the quantile lines from each method should be nearly equivalent on data with low contamination.

Also from the equivalence of the coefficient estimates, the interpretation of the L_2E coefficients will be the same as the interpretation provided by Koenker and Hallock. In particular, we note that the coefficients estimates for most of the predictor variables are outside the confidence interval of the least squares coefficient estimates. This means that the effects these variables have on the conditional distribution are not constant across all quantiles, and thus the quantile estimates are influenced by more than just the intercept term. For example, given all other factors remain the same, a boy in the 10th percentile would be expected to be about .05 kg larger than a girl in the 10th percentile, while a boy in the 90th percentile would be expected to be about .14 kg larger than a girl in the 90th percentile. Using least squares regression to estimate the conditional quantiles would not capture this difference in effects.

6.2 Personal Income Data

For this study, we are attempting to estimate the quantiles of the conditional distribution of personal income given age. The data set used is a subset of the 1995 Current Population Survey, conducted by the Bureau of the Census for the Bureau of Labor Statistics. In particular, this subset consists of people with at least a bachelor's degree, who work full time (at least 40 hours a week), who have a non-negative reported personal income, and who live in the northeastern United States. With these restrictions, the subset contains 3383 observations. The histograms of the variables

Age, Personal Income, and Log Personal Income can be found in Figure 6.2. This study focuses on two variables: Age and Log Personal Income.



(a) Age (b) Income (c) Log Income

Figure 6.2 : Histograms for the age and both the regular and logged personal income variables.

We first attempt to use simple linear quantile regression to obtain our quantile estimates. The resulting quantile lines, for both L_2E quantile regression and KB's quantile regression, can be found in Figure 6.3. We see that the L_2E regression lines for the lower quantiles are closer to the median than the classical quantile regression lines, leading us to believe that the lower outliers are affecting the coefficient estimates using KB's quantile regression. A difference between the coefficient estimates from each method can be seen in Figure 6.4. Using our L_2E coefficient estimates, we can then give estimates for the distributions of personal income for a given age. Estimates for the 025, .05, .1, .25, .5, .75, .9, .95, and .975 quantiles for each age in the range $\{25, 30, \ldots, 65, 70\}$ can be found in Table 6.1.

When examining the plot of Age against Log Personal Income, we see that there appears to be a positive trend from roughly ages 20 to 35, followed by a period with no trend, and then a slightly negative trend from about age 55 and beyond. These changing trends indicate that a linear fit is probably not appropriate for this data.

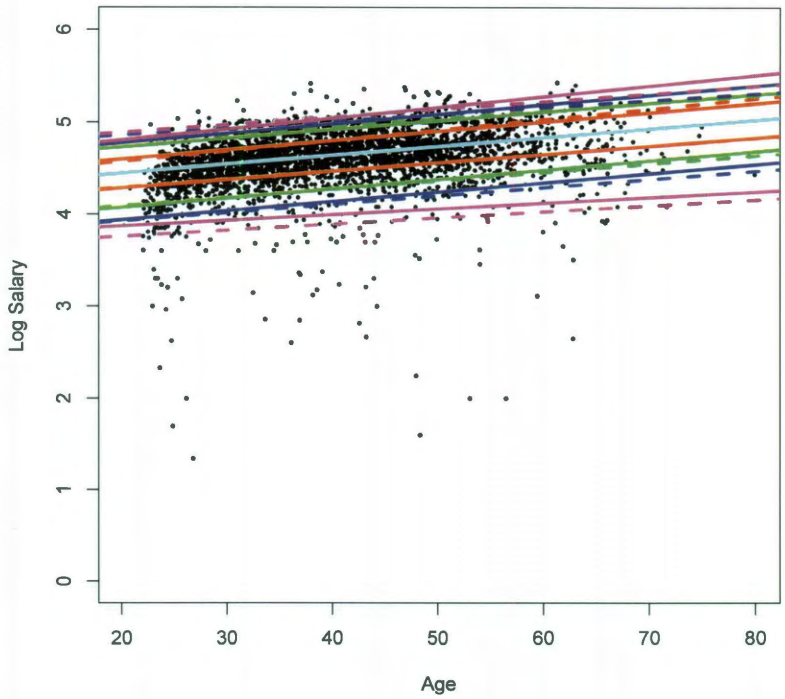


Figure 6.3 : Plot of age against log personal income with simple linear quantile regression lines. L_2E regression lines are solid while KB's regression lines are dashed. The lines shown are the .025, .05, .1, .25, .5, .75, .9, .95, and .975 quantile estimates.

So, we first attempt to use L_2E quantile regression with cubic splines to obtain both a better idea of the trend of the data as well as a better estimate of the conditional quantiles. To do this, knots were set at ages 35 and 55, that is, the ages where we think that the trends may change. From this, we obtain the curves seen in Figure 6.5. The conditional quantile estimates derived from these curves can be fond in Table 6.2.

One of the most noticeable things about Figure 6.5 is the early positive trend in

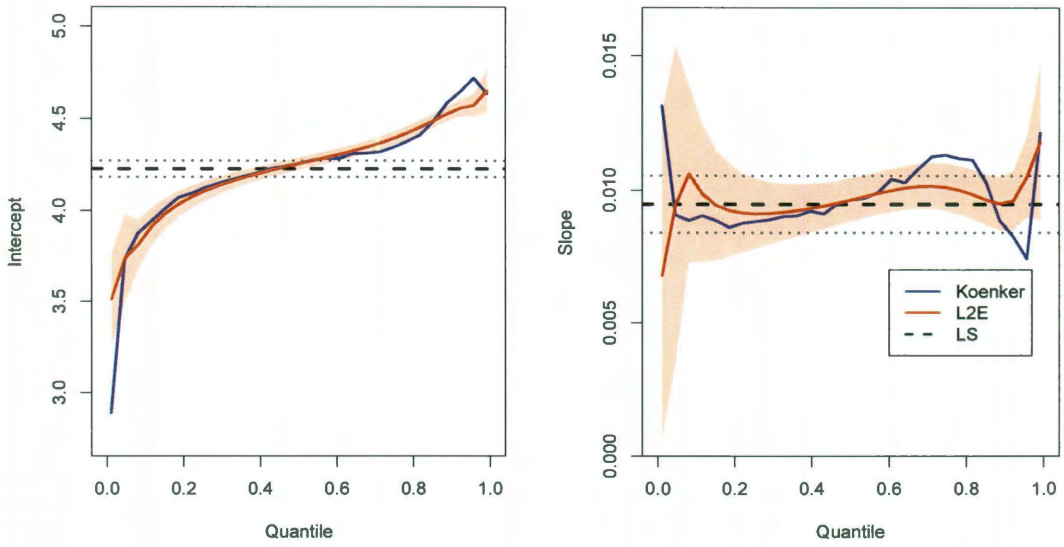


Figure 6.4 : Comparison of coefficient estimates from L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the relationship between age and log personal income. The shaded red regions represent the 95% confidence intervals for the L_2E quantile regression coefficient estimates. The least squares coefficient estimates, along with their 95% confidence interval, are shown in black.

	25	30	35	40	45	50	55	60	65	70
0.025	7.986	8.581	9.219	9.905	10.642	11.434	12.285	13.199	14.182	15.237
0.05	9.768	10.969	12.317	13.831	15.532	17.441	19.585	21.993	24.697	27.733
0.1	13.345	14.999	16.858	18.948	21.297	23.937	26.904	30.239	33.987	38.200
0.25	21.423	23.793	26.425	29.348	32.595	36.201	40.205	44.653	49.592	55.078
0.5	31.338	35.012	39.117	43.703	48.826	54.550	60.946	68.091	76.073	84.992
0.75	44.555	50.048	56.219	63.150	70.935	79.681	89.505	100.540	112.935	126.859
0.9	60.008	66.937	74.666	83.287	92.903	103.630	115.596	128.943	143.831	160.439
0.95	67.489	75.973	85.524	96.275	108.378	122.002	137.339	154.604	174.039	195.918
0.975	73.916	84.527	96.662	110.538	126.407	144.554	165.307	189.038	216.176	247.210

Table 6.1 : Conditional quantile estimates for various ages derived from simple linear quantile regression on log personal income, displayed in thousands of dollars.

all of the conditional quantile curves, as we conjectured. This trend continues until roughly the age of 35, where the curves level out for several years. During these years, the conditional distributions given Age are close to equivalent. Then, starting around age 50, we see the conditional distributions begin to widen, as the upper quantile curves have a positive trend while the lower quantile curves have a slightly negative trend. The median quantile regression curve, however, exhibits very little trend, as evidenced by the small change in conditional median estimates in the table for ages 45 to 70.

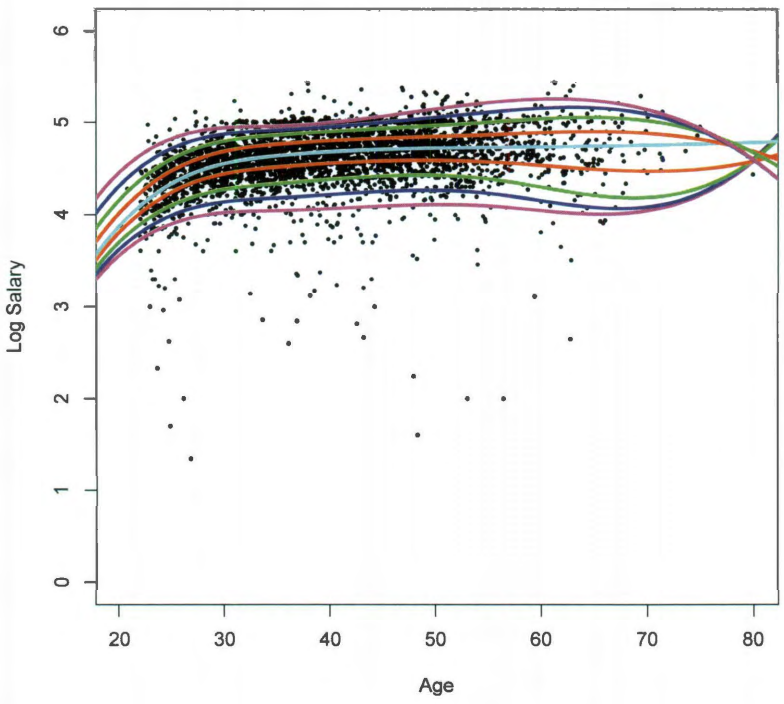


Figure 6.5 : L_2E quantile regression with cubic splines for log personal income. The curves shown are the .025, .05, .1, .25, .5, .75, .9, .95, and .975 quantile estimates.

	25	30	35	40	45	50	55	60	65	70
0.025	7.916	10.561	11.154	11.723	12.446	12.840	12.385	11.066	10.206	11.151
0.05	9.712	13.906	15.516	16.812	17.937	18.150	16.776	14.078	12.094	12.371
0.1	12.445	18.773	21.830	24.285	26.223	26.513	24.214	19.851	16.348	15.592
0.25	17.485	27.601	32.705	36.026	38.254	38.836	37.391	34.214	31.079	29.578
0.5	21.577	35.592	43.276	47.888	50.855	52.539	53.529	54.422	55.456	56.786
0.75	30.067	45.593	58.311	65.618	68.248	69.039	71.477	77.815	83.192	79.643
0.9	39.978	62.699	72.996	79.640	85.945	92.597	100.529	109.582	113.501	104.364
0.95	52.361	75.123	82.483	89.603	100.223	113.782	129.240	143.234	144.701	123.158
0.975	66.908	87.632	91.122	99.647	116.834	140.243	164.572	179.903	173.139	138.353

Table 6.2 : Conditional quantile estimates for various ages derived from quantile regression with cubic splines on log personal income, displayed in thousands of dollars.

As an alternative to cubic splines, we also performed L_2E local linear quantile regression on the log personal income data. Using a gaussian kernel, we used different bandwidths for the various quantiles. For the most extreme quantiles, that is, the .025 and .975 quantiles, we used a bandwidth of 20. For the .05 and .95 quantiles, we used a bandwidth of 15. For the other five quantiles, we used a bandwidth of 5. This was done to improve the stability of the most extreme quantile estimates. The resulting curves can be seen in Figure 6.6. Table 6.3 shows the estimated quantiles of personal income given various ages.

Once again, we see a positive trend in the quantile curves until sometime between ages 35 and 40 followed by a period of no trend. We do see a more pronounced downward trend past age 65, but it is unclear whether that is a real trend, or if the trend is caused by the lack of data in that region. The local linear quantile curves do give similar results as the quantile curves derived using cubic splines, particularly at the median, .1, and .9 quantile levels, validating our results from the cubic splines. Because of the relative levels of stability, we find it preferable to use the results from the quantile curves derived using cubic splines, although the results from the local linear quantile curves could be used as well.

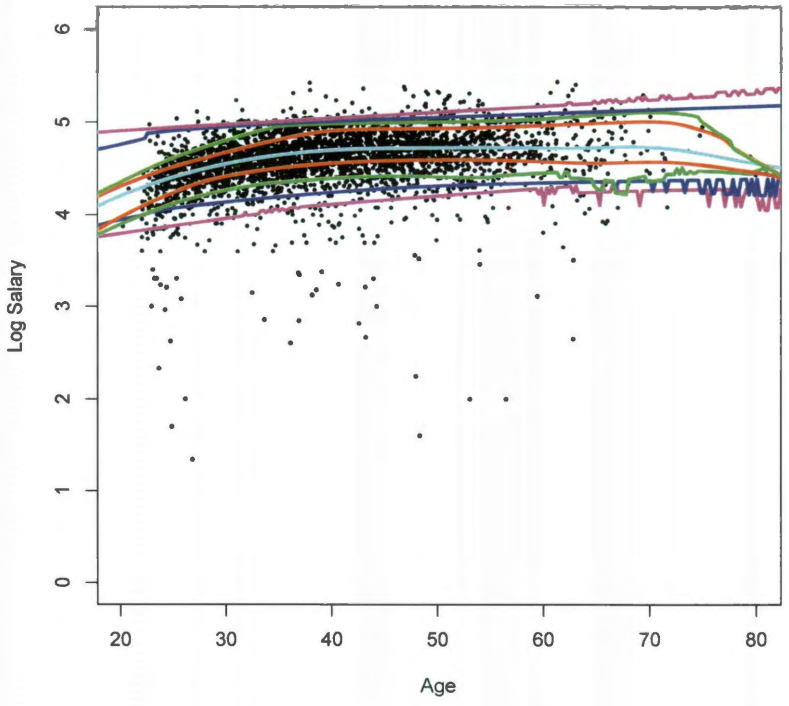


Figure 6.6 : Estimated quantile values for various theoretic quantiles of L_2E local linear quantile regression on log personal income. The curves shown are the .025, .05, .1, .25, .5, .75, .9, .95, and .975 quantile estimates.

	25	30	35	40	45	50	55	60	65	70
0.025	7.604	9.173	11.015	13.014	14.473	16.020	17.682	16.879	21.236	18.383
0.05	11.780	13.801	15.700	17.570	19.227	20.602	21.589	22.254	22.909	23.482
0.1	13.220	18.467	22.351	24.825	26.266	25.830	25.555	28.460	19.295	23.098
0.25	17.397	25.968	31.831	35.975	38.315	39.358	39.121	37.442	36.230	37.223
0.5	25.459	35.333	44.019	49.895	52.319	52.943	53.499	52.736	53.323	53.714
0.75	32.690	47.093	64.604	80.538	84.521	85.600	87.817	92.189	97.131	101.258
0.9	41.013	63.488	90.180	95.727	98.059	100.517	103.306	108.901	119.068	127.301
0.95	80.783	89.366	96.477	102.515	107.655	112.697	117.945	123.497	129.383	135.595
0.975	87.537	94.736	102.254	110.222	118.740	127.893	137.766	148.448	160.042	172.679

Table 6.3 : Conditional quantile estimates for various ages derived from local linear quantile regression on log personal income, displayed in thousands of dollars.

Intuitively, the trends that these non-linear quantile curves show make a great deal of sense and would not have been seen using linear quantile regression. From both the non-linear methods, we see that there is initially an upward trend from ages 20 to roughly 35, followed by a plateau from ages 35 to 55, and then a slight decline from age 55 to beyond. The first trend likely follows from people entering the workforce and receiving raises as they begin their career. The middle section probably stems from people hitting their career peak along with other factors such as changing careers, firings, and the like. The final section likely relates to the declining workforce, as seen in Figure 6.2(a). Although mortality is a consideration to the declining workforce, it seems unlikely to be the sole cause of the declining income estimates. Instead, it is likely that what is happening is that people who have a higher income are able to retire earlier, while those with a lower income have to continue working. We also notice that the spread of the quantiles estimates tends to increase as age increases, at least until we get to the upper range of age. This is due to the upper quantile estimates increasing as age increase, as the lower quantiles stay relatively constant. This leads us to believe that incomes in the upper quantiles increase at a faster rate than the incomes in the lower quantiles.

6.3 Baseball Player Valuation

For this study, Major League Baseball (MLB) player-seasons from the years 2001 to 2010 are examined to determine the estimated quantiles for the market value of individual players. That is, we want to be able to estimate what we expect the market value of a player to be based on his statistics and the value that the market places on those statistics. The median estimate is treated as that expected value, but the other quantile levels are estimated as well, giving us what we will call a value distribution.

This gives us an idea of a reasonable range of salaries for a player based on his performance. We study both position players and pitchers. The data is taken from The Lahman Baseball Database, which can be found at http://baseball1.com/statistics/.

6.3.1 Position Players

To examine position players, we first limit our sample to player-seasons in which the player accumulated at least 300 at-bats and had a salary above a set minimum, in this case $414,000 (slightly above the MLB minimum).This leaves us with 1771 observations, with16 predictor variables, and one response variable (Salary). The predictor variables are number of at-bats (AB), runs scored (R), number of singles, doubles, triples and home runs hit (H,X2B, X3B, and HR), runs batted in (RBI), stolen bases (SB), times caught stealing (CS), number of walks (BB), number of strikeouts (SO), number of intentional walks (IBB), times hit by a pitch (HBP), number of sacrifice hits (SH), number of sacrifice flies (SF), and times grounded into a double play (GIDP). For each year, the data has been standardized (that is, for example, a batter who hit 19.98 home runs in 2001 would have the same standardized value as a batter who hit 16.59 home runs in 2005, as those were the mean numbers of home runs for their respective years). The full linear model, can be written as

$$\begin{aligned}\text{Salary} =& \beta_0 + \beta_1\text{AB} + \beta_2\text{R} + \beta_3\text{H} + \beta_4\text{X2B} + \beta_5\text{X3B} + \beta_6\text{HR} + \beta_7\text{RBI} + \beta_8\text{SB} + \beta_9\text{CS} \\ &+ \beta_{10}\text{BB} + \beta_{11}\text{SO} + \beta_{12}\text{IBB} + \beta_{13}\text{HBP} + \beta_{14}\text{SH} + \beta_{15}\text{SF} + \beta_{16}\text{GIDP} + \epsilon.\end{aligned}$$

The coefficient estimates from this model using both L_2E and KB's quantile regression can be seen in Figure 6.7. As we can see, there aren't that many differences between the two sets of coefficient estimates, leading us to believe that there aren't many

outliers in this data. We can then use these coefficient plots to estimate the value distributions of individual players.

Believing there may be some collinearity among the variables, we attempt to reduce the model using the AIC criteria described in Section ??. A best-subsets analysis gives us the results in Table 6.4. As we can see, the model with the lowest AIC value is the 7-factor model given as

$$\text{Salary} = \beta_0 + \beta_1 \text{R} + \beta_2 \text{X2B} + \beta_3 \text{X3B} + \beta_4 \text{CS} + \beta_5 \text{BB} + \beta_6 \text{HBP} + \beta_7 \text{SH} + \epsilon.$$

It is worth noting that this is somewhat different from the reduced model selected for least-squares regression, which is given by

$$\begin{aligned}\text{Salary} = &\beta_0 + \beta_1 \text{R} + \beta_2 \text{X2B} + \beta_3 \text{X3B} + \beta_4 \text{RBI} + \beta_5 \text{BB} + \beta_6 \text{SO} + \beta_7 \text{IBB} \\ &+ \beta_6 \text{SH} + \beta_7 \text{GIDP} + \epsilon.\end{aligned}$$

The coefficient estimates from our reduced model using both forms of quantile regression can be seen in Figure 6.8. Once again, there is not much difference in the coefficient estimates from using each method.

The median estimates for each player in the position player data set from the year 2010 can be found in Appendix B.1. The estimates using both the full and AIC-reduced models are shown, along with the difference between those estimates and the actual salary of the individual player. From the reduced model, we see that the most overpaid player was Alex Rodriguez, with a difference of -26.446 million dollars, and that the most underpaid player was Joey Votto, with a difference of 9.677 million dollars. Likewise, we see that the player with the highest value was Prince Fielder,

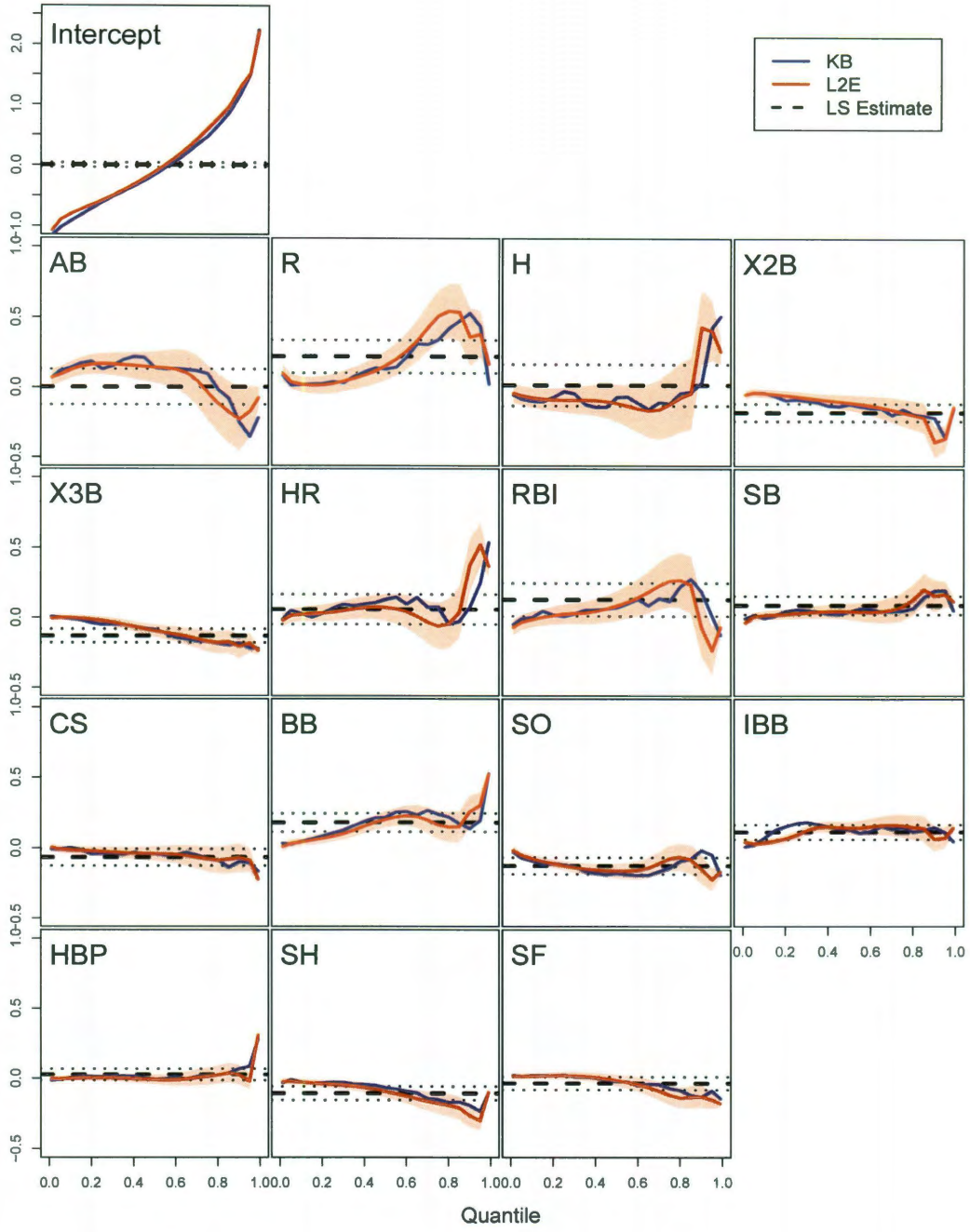


Figure 6.7 : Comparison of coefficient estimates from L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the position player data set. The shaded red regions represent the 95% confidence intervals for the L_2E quantile regression coefficient estimates. The least squares coefficient estimates, along with their 95% confidence interval, are shown in black.

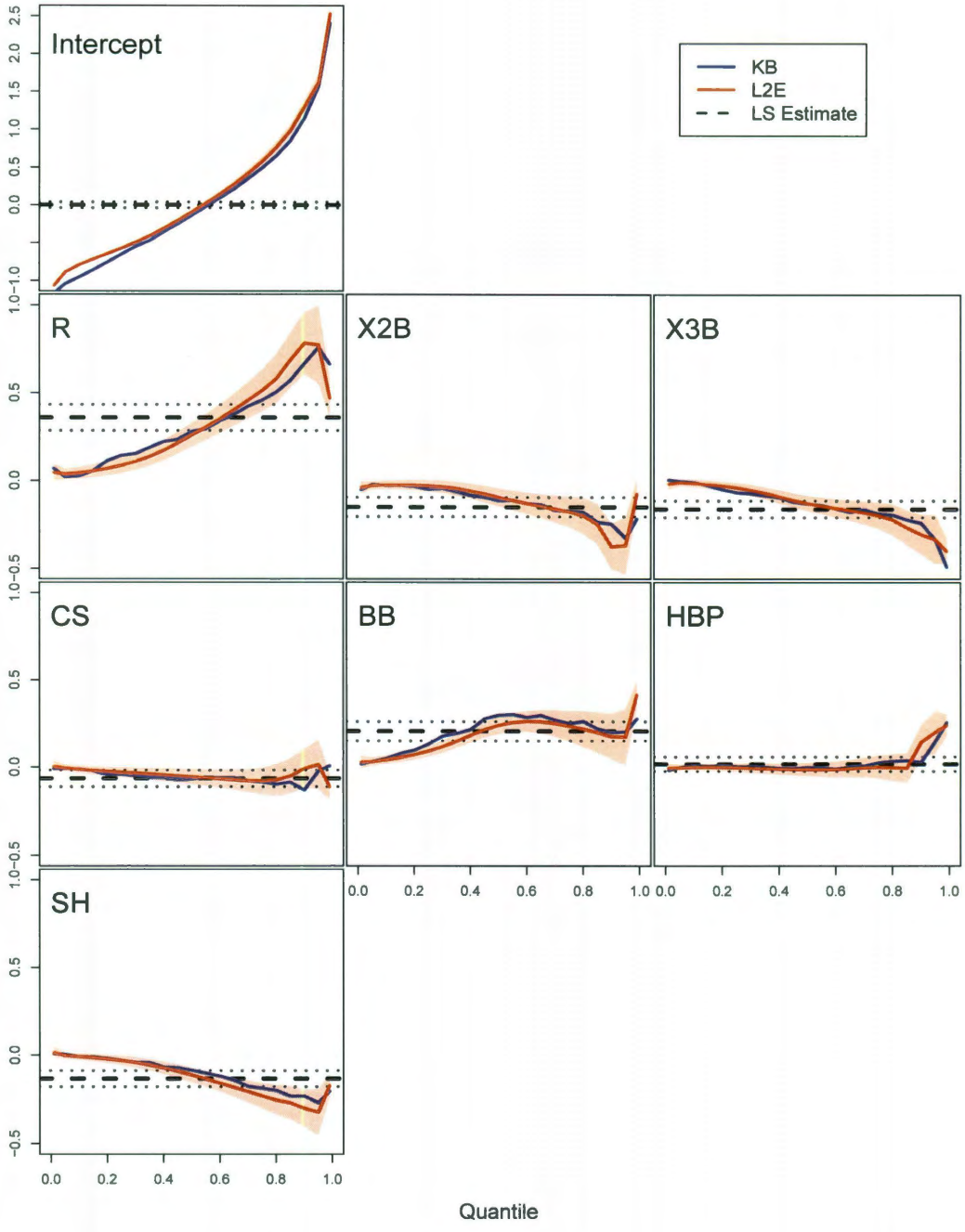


Figure 6.8 : Comparison of coefficient estimates from L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the reduced position player data set, using AIC as the selection criteria. The shaded red regions represent the 95% confidence intervals for the L_2E quantile regression coefficient estimates. The least squares coefficient estimates, along with their 95% confidence interval, are shown in black.

p	AB	R	H	X2B	X3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP	AIC
1	F	F	F	F	F	F	F	F	F	T	F	F	F	F	F	F	-565.39
2	F	F	F	F	F	F	F	F	F	T	F	F	F	T	F	F	-574.529
3	F	F	F	F	T	F	F	F	F	T	F	F	F	T	F	F	-575.524
4	F	F	F	F	T	F	F	F	T	T	F	F	F	T	F	F	-575.594
5	F	T	F	T	F	F	F	T	F	T	F	F	F	T	F	F	-575.541
6	F	T	F	T	T	F	F	F	T	T	F	F	F	T	F	F	-576.064
7	F	T	F	T	T	F	F	F	T	T	F	F	T	T	F	F	-576.131
8	F	T	F	T	T	F	F	T	T	T	F	F	T	T	F	F	-575.847
9	F	T	F	T	T	T	T	F	T	T	F	F	T	T	F	F	-574.38
10	F	T	F	T	T	T	T	T	T	T	F	F	T	T	F	F	-573.248
11	F	T	F	T	T	T	T	T	T	T	F	F	T	T	F	T	-570.964
12	F	T	F	T	T	T	T	T	T	T	F	F	T	T	T	T	-567.808
13	T	T	T	T	T	T	T	T	T	T	F	F	T	T	F	T	-563.549
14	T	T	T	T	T	T	T	T	T	T	F	F	T	T	T	T	-559.119
15	T	T	T	T	T	T	T	T	T	T	T	F	T	T	T	T	-536.372
16	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-514.034

Table 6.4 : Best linear models fitted to position player data using AIC as the selection criterion for each value of p, the number of factors in the model.

with a value of 12.306 million dollars, while the player with the lowest value was Nyjer Morgan, who actually had a negative value of -0.594 million dollars.

6.3.2 Pitchers

Similar to the position players, we examined MLB player-seasons for pitchers. Each of the player-seasons included in this study are such that the player appeared in at least 15 games and made more the the same imposed minimum salary as the batters, that is $414,000. The data set includes 2223 observations and 15 predictor variables and one response variable (Salary). The predictor variables are the number of wins (W), losses (L), games started (GS), games relieved (GR), complete games (CG), saves (SV), number of outs pitched (IPouts), hits allowed (H), earned runs allowed (ER), home runs allowed (HR), walks allowed (BB), number of strike outs (SO), number of intentional walks (IBB), number of batters hit by a pitch (HBP), and number of times the pitcher was called for a balk (BK). The variables H, ER, HR, BB, SO, IBB,

HBP, and BK are the number of each variable the pitcher averages per nine innings. That is, if the pitcher gave up 350 earned runs in 100 innings, the value of their ER variable would be 3.5. As before with the batters, the data has been standardized by year. The full model can be written as

$$\begin{aligned}\text{Salary} =&\beta_0 + \beta_1\text{W} + \beta_2\text{L} + \beta_3\text{GS} + \beta_4\text{GR} + \beta_5\text{CG} + \beta_6\text{SV} + \beta_7\text{IPouts} + \beta_8\text{H} + \beta_9\text{ER} \\ &+ \beta_{10}\text{HR} + \beta_{11}\text{BB} + \beta_{12}\text{SO} + \beta_{13}\text{IBB} + \beta_{14}\text{HBP} + \beta_{15}\text{BK} + \epsilon.\end{aligned}$$

Using this model gives us the coefficient plots found in Figure 6.9. Although it is possible to use these estimates to create the value distribution of individual players, it is reasonable to assume that there are actually two groups in this model, starting pitchers and relief pitchers. Thus, we divide the pitcher data set into two subsets based on if the player is a starting pitcher or a relief pitcher. To do so, we define a relief pitcher to be a pitcher with twice as many games relieved as games started. That is, GR $>$ 2 GS. This leaves us with 997 starting pitchers and 1226 relief pitchers.

For the starting pitchers data set, we remove the reliever-specific predictor variables, namely SV and GR, which gives us the full model

$$\begin{aligned}\text{Salary} =&\beta_0 + \beta_1\text{W} + \beta_2\text{L} + \beta_3\text{GS} + \beta_4\text{CG} + \beta_5\text{IPouts} + \beta_6\text{H} + \beta_7\text{ER} \\ &+ \beta_8\text{HR} + \beta_9\text{BB} + \beta_{10}\text{SO} + \beta_{11}\text{IBB} + \beta_{12}\text{HBP} + \beta_{13}\text{BK} + \epsilon.\end{aligned}$$

The coefficient estimates can be seen in Figure 6.10. In these plots, we see some slight differences in the estimates from the two quantile regression methods. In particular, wins are slightly more valued by L_2E than KB's Method, while number of games started are slightly less valued by L_2E. However, for the most part, the estimates

from KB's method fall within the confidence interval of the L_2E coefficients.

We again attempt to reduce this model using AIC as our selection criteria. The results of the best-subsets analysis can be found in Table 6.5. As we can see, the model with the lowest calculated AIC value is the 7-factor model given by

$$\text{Salary} = \beta_0 + \beta_1 \text{L} + \beta_2 \text{H} + \beta_3 \text{ER} + \beta_4 \text{HR} + \beta_5 \text{BB} + \beta_6 \text{IBB} + \beta_7 \text{HBP} + \epsilon.$$

This is different than the reduced model selected using least-squares regression, which is given by

$$\text{Salary} = \beta_0 + \beta_1 \text{W} + \beta_2 \text{IPouts} + \beta_3 \text{BB} + \beta_4 \text{SO} + \epsilon.$$

The coefficient estimates from our reduced model using both forms of quantile regression can be seen in Figure 6.11. Unlike the full model, there is not much difference in the coefficient estimates from using each method.

p	W	L	GS	CG	IPouts	H	ER	HR	BB	SO	IBB	HBP	BK	AIC
1	F	F	F	F	F	F	F	F	T	F	F	F	F	-319.775
2	F	T	F	F	F	F	F	F	T	F	F	F	F	-320.395
3	F	T	F	F	F	F	F	F	T	F	T	F	F	-320.957
4	F	T	F	F	F	T	F	F	T	F	T	F	F	-321.164
5	F	T	F	F	F	T	F	F	T	F	T	T	F	-321.37
6	F	T	F	F	F	T	T	F	T	F	T	T	F	-321.526
7	F	T	F	F	F	T	T	T	T	F	T	T	F	-321.513
8	F	T	F	F	F	T	T	T	T	T	T	T	F	-320.977
9	F	T	F	F	F	T	T	T	T	T	T	T	T	-317.002
10	T	T	T	T	T	T	F	F	T	T	T	T	F	-314.183
11	T	T	T	T	T	T	F	T	T	T	T	T	F	-313.193
12	T	T	T	T	T	T	T	T	T	T	F	T	T	-311.604
13	T	T	T	T	T	T	T	T	T	T	T	T	T	-310.229

Table 6.5 : Best linear models fitted to starting pitcher data using AIC as the selection criterion for each value of p, the number of factors in the model.

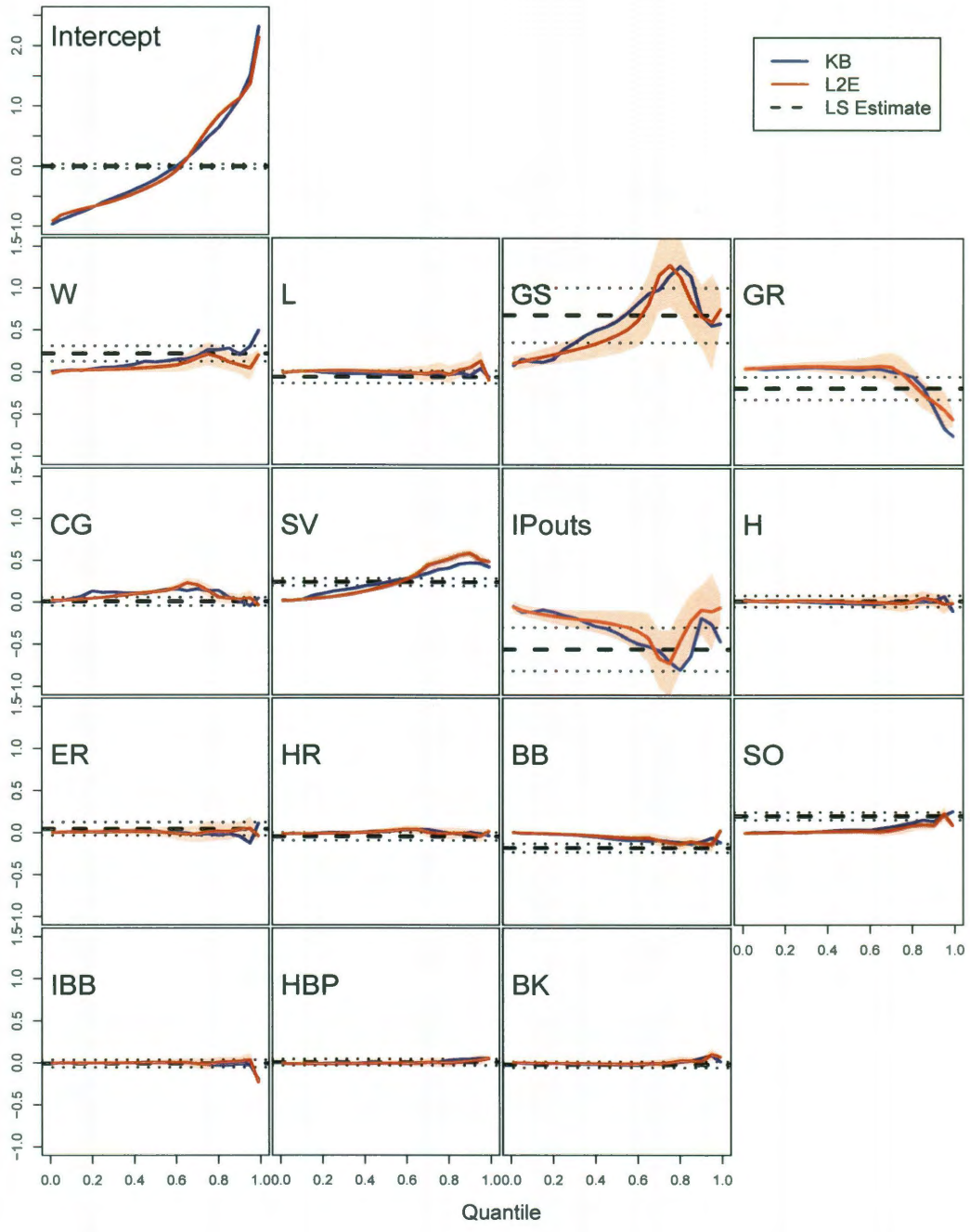


Figure 6.9 : Comparison of coefficient estimates from L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the full pitcher data set. The shaded red regions represent the 95% confidence intervals for the L_2E quantile regression coefficient estimates. The least squares coefficient estimates, along with their 95% confidence interval, are shown in black.

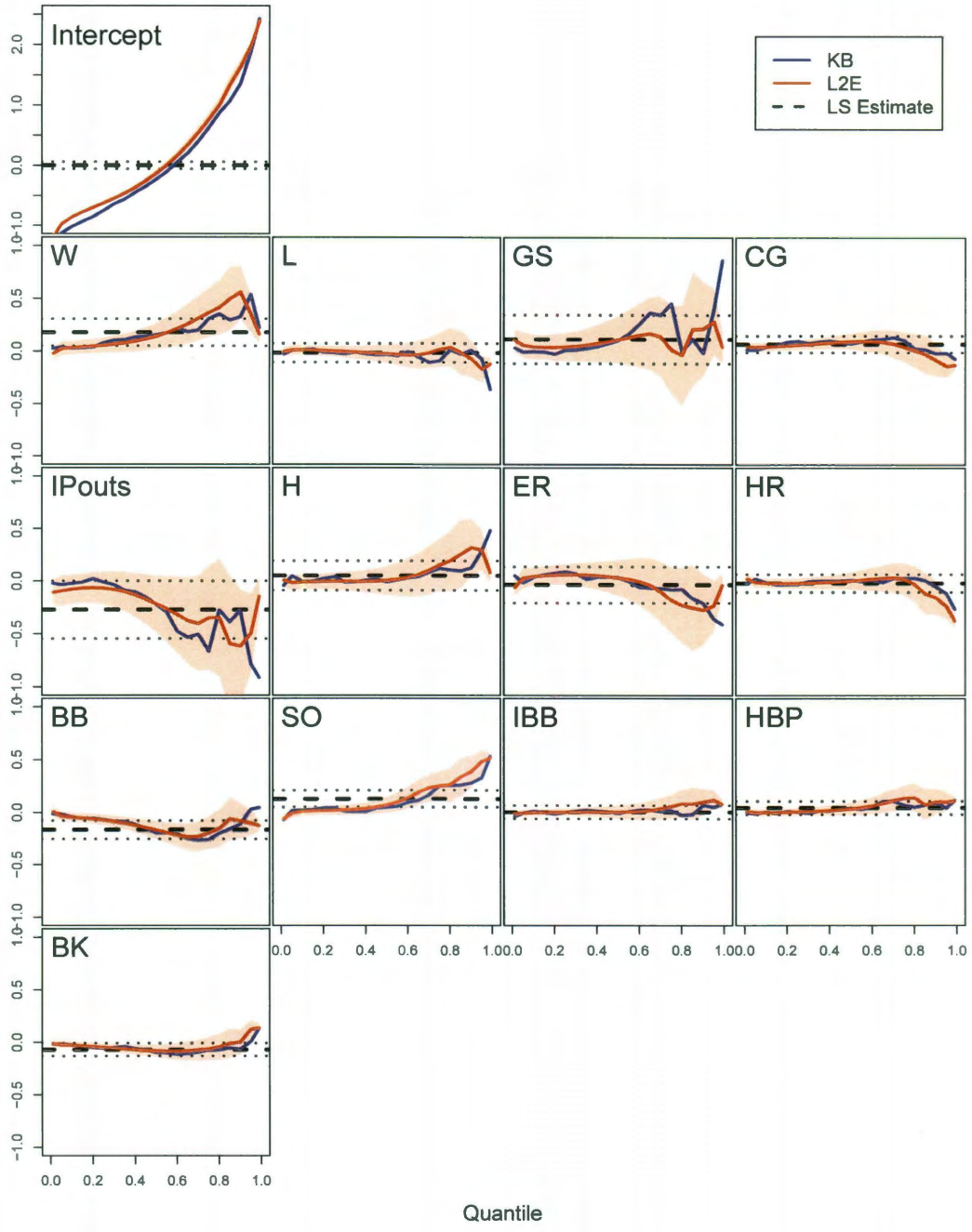


Figure 6.10 : Comparison of coefficient estimates from L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the starting pitcher data set. The shaded red regions represent the 95% confidence intervals for the L_2E quantile regression coefficient estimates. The least squares coefficient estimates, along with their 95% confidence interval, are shown in black.

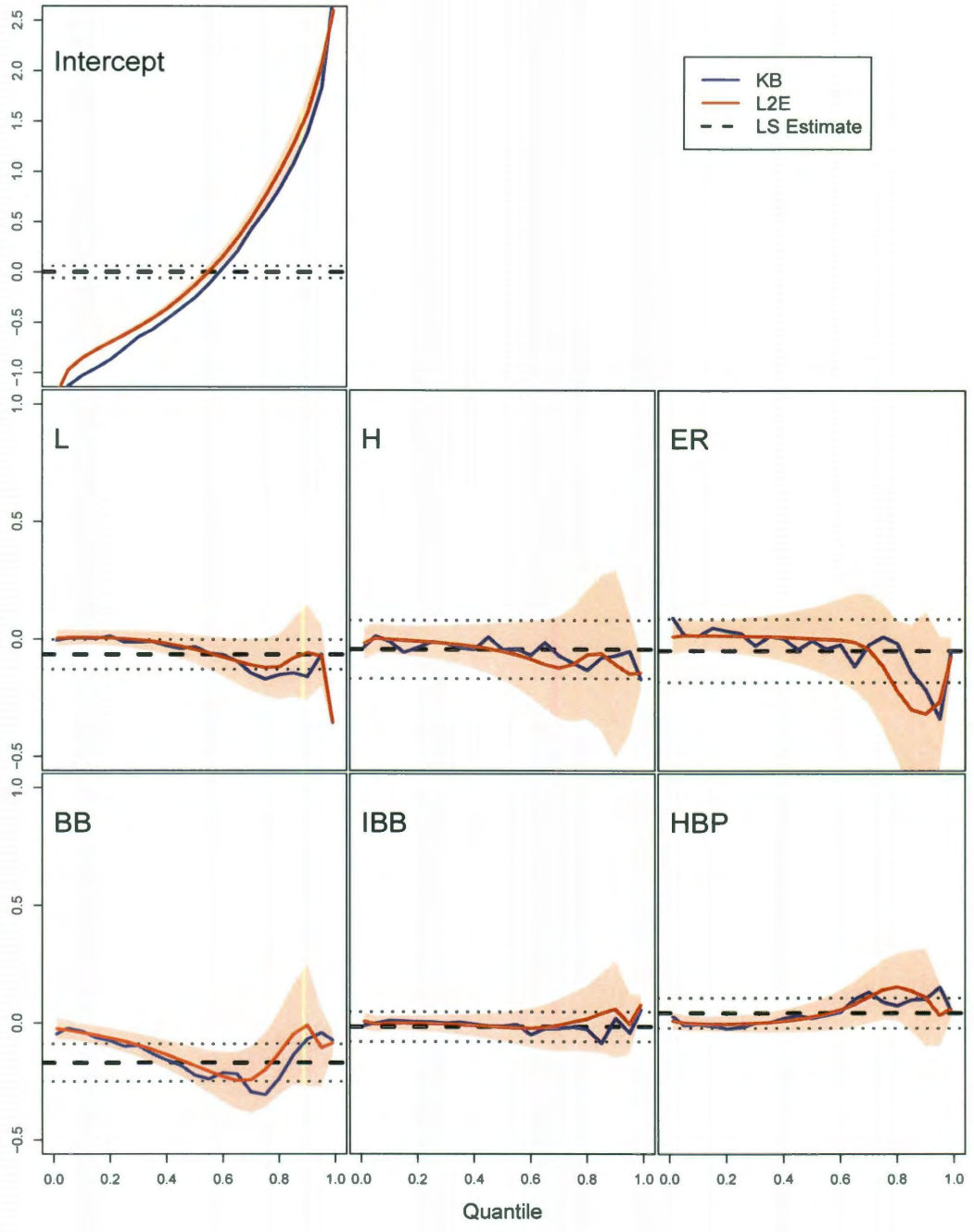


Figure 6.11 : Comparison of coefficient estimates from L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the reduced starting pitcher data set, using AIC as the selection criteria. The shaded red regions represent the 95% confidence intervals for the L_2E quantile regression coefficient estimates. The least squares coefficient estimates, along with their 95% confidence interval, are shown in black.

The median estimates for each player in the starting pitcher data set from the year 2010 can be found in Appendix B.2.1. The estimates using both the full and AIC-reduced models are shown, along with the difference between those estimates and the actual salary of the individual player. From the reduced model, we see that the most overpaid player was CC Sabathia, with a difference of -18.327 million dollars, and that the most underpaid player was Kevin Slowey, with a difference of 6.430 million dollars. Cliff Lee was the starting pitcher with the highest estimated value, with a value of 7.762 million dollars, while the player with the lowest value was Oliver Perez, who had an estimated value of 0.194 million dollars.

Once again we repeat this process using the relief pitcher data set. For this model, we remove the starter-specific variables, namely GS and CG, from our full pitcher linear model, leaving us with full relief pitcher model given by

$$\begin{aligned}\text{Salary} =&\beta_0 + \beta_1\text{W} + \beta_2\text{L} + \beta_3\text{GR} + \beta_4\text{SV} + \beta_5\text{IPouts} + \beta_6\text{H} + \beta_7\text{ER} \\ &+ \beta_8\text{HR} + \beta_9\text{BB} + \beta_{10}\text{SO} + \beta_{11}\text{IBB} + \beta_{12}\text{HBP} + \beta_{13}\text{BK} + \epsilon.\end{aligned}$$

The coefficient estimates from using both methods of quantile regression can be found in Figure 6.12. Although many of the coefficient estimate plots are similar, we see a noticeable difference in the coefficient estimates for SV. In particular, L_2E places less value on saves. This seems likely due to the importance in general placed on the idea of closers. This small group of relief pitchers is paid a much higher salary than the other relievers. L_2E quantile regression would then treat these points as outliers, making these points less influential on the coefficient estimates.

We attempt find a reduced model using a best-subsets analysis with AIC as our selection criteria. The results of this analysis can be found in Table 6.6. As we can

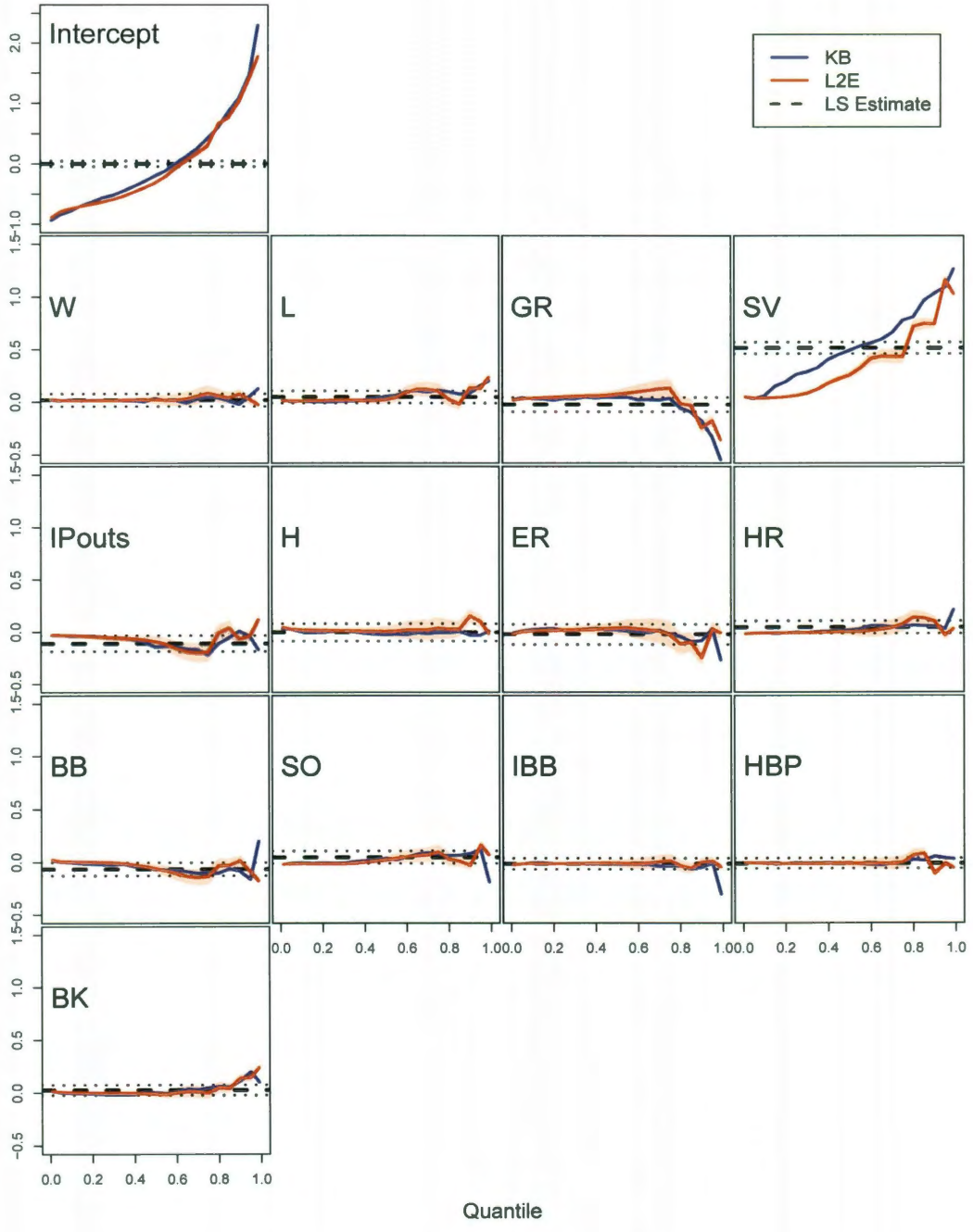


Figure 6.12 : Comparison of coefficient estimates from L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the relief pitcher data set. The shaded red regions represent the 95% confidence intervals for the L_2E quantile regression coefficient estimates. The least squares coefficient estimates, along with their 95% confidence interval, are shown in black.

see, the model with the lowest AIC value is the 5 factor model given by

$$\text{Salary} = \beta_0 + \beta_1 \text{L} + \beta_2 \text{IPouts} + \beta_3 \text{BB} + \beta_4 \text{SO} + \beta_5 \text{HBP} + \epsilon.$$

Unlike the previous reduced models, this is noticeably different than the reduced model chosen for least-squares regression, given by

$$\text{Salary} = \beta_0 + \beta_1 \text{SV} + \beta_2 \text{IPouts} + \epsilon.$$

The coefficients estimates from our reduced model can be found in Figure 6.13. We still see a difference in the coefficient estimates from the two methods, most clearly in the plot of SO. This means that we will see a difference in the value distributions estimated by the two methods.

p	W	L	GR	SV	IPouts	H	ER	HR	BB	SO	IBB	HBP	BK	AIC
1	F	T	F	F	F	F	F	F	F	F	F	F	F	-153.574
2	F	T	F	F	F	F	F	F	F	F	T	F	F	-156.316
3	F	T	F	F	T	F	F	F	T	F	F	F	F	-160.059
4	F	T	F	F	T	F	F	F	T	T	F	F	F	-166.955
5	F	T	F	F	T	F	F	F	T	T	F	T	F	-167.453
6	F	T	F	F	T	F	F	F	T	T	T	T	F	-166.595
7	F	T	F	F	T	T	F	F	T	T	T	T	F	-164.476
8	F	T	F	F	T	T	F	F	T	T	T	T	T	-162.559
9	F	T	F	F	T	T	T	F	T	T	T	T	T	-160.859
10	F	T	F	F	T	T	T	T	T	T	T	T	T	-158.709
11	T	T	F	F	T	T	T	T	T	T	T	T	T	-154.095
12	T	T	F	T	T	T	T	T	T	T	T	T	T	-144.238
13	T	T	T	T	T	T	T	T	T	T	T	T	T	-140.52

Table 6.6 : Best linear models fitted to relief pitcher data using AIC as the selection criterion for each value of p, the number of factors in the model.

The median estimates for each player in the relief pitcher data set from the year 2010 can be found in Appendix B.2.2. The estimates using both the full and AIC-

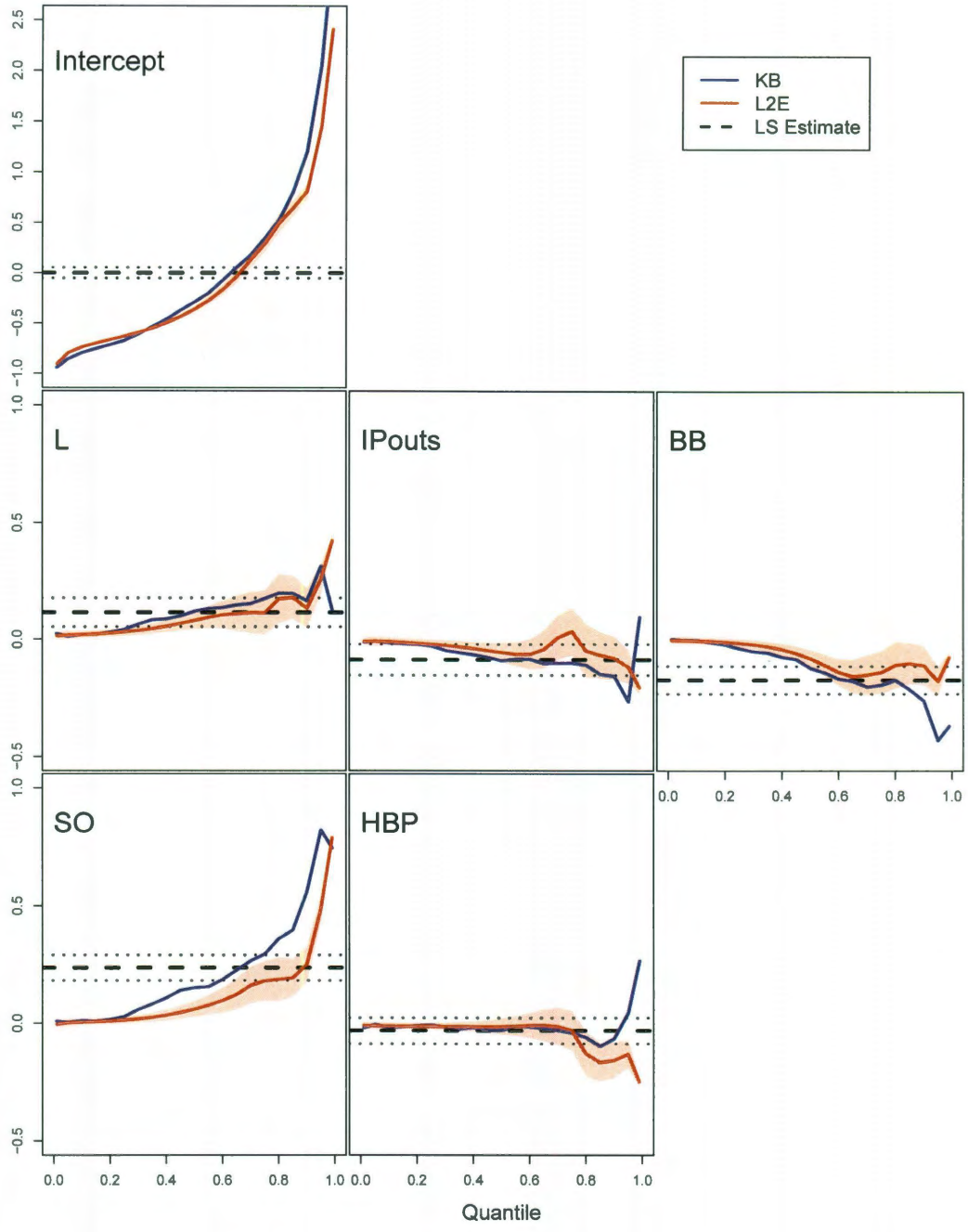


Figure 6.13 : Comparison of coefficient estimates from L_2E quantile regression, shown in red, and KB's quantile regression, shown in blue, on the reduced relief pitcher data set, using AIC as the selection criteria. The shaded red regions represent the 95% confidence intervals for the L_2E quantile regression coefficient estimates. The least squares coefficient estimates, along with their 95% confidence interval, are shown in black.

reduced models are shown, along with the difference between those estimates and the actual salary of the individual player. From the reduced model, we see that the most overpaid player was Mariano Rivera, with a difference of -11.917 million dollars, and that the most underpaid player was Edward Mujica, with a difference of 2.803 million dollars. Mujica was also the relief pitcher with the highest estimated value, with a value of 3.222 million dollars, while the player with the lowest value was Brian Bruney, who also had a negative estimated value of -0.290 million dollars.

6.3.3 Arbitration Results

After a player in MLB has been in the league for three years, he then becomes eligible for arbitration, a process that determines the player's salary. Before this time, the player's salary is determined by the team and is usually near the league minimum salary. In the arbitration process, a player submits what they feel their salary should be and the team submits what they feel the player's salary should be. In most instances, the team and the player agree to a contract that is somewhere between the two salary values, but in some cases, the case goes to arbitration. When this happens, an arbiter determines which of the two submitted salary numbers represents the player's value best. The chosen number then becomes the player's salary for the following year. This is a good opportunity to compare our L_2E median estimates to a real world valuation of players.

After the 2010 season, only 3 players had their salaries determined by an arbiter. The one position player whose salary was determined by an arbiter was Hunter Pence. He submitted a salary of \$6.9MM while the team submitted a salary of \$5.15MM, with the median estimate from the L_2E reduced model being in-between the two at \$5.905MM. Pence won the hearing, although the team's submission was slightly closer

to the model's estimated value. The two pitchers who had their salary determined by an arbiter were both starting pitchers. Ross Ohlendorf won his arbitration hearing, giving him a salary of $2.025, which was higher than the team's offer of $1.4MM, but still lower than the estimated salary of $4.772MM. Jared Weaver lost his arbitration hearing, giving him the salary of $7.375MM, which was lower than his submission of $8.8MM, but still higher than the estimated salary of $6.228MM. In both of these cases, the arbiter ruled in the direction of the model's estimated value.

6.3.4 Conclusions

One of the first things we notice about the results from our three reduced models is that player salaries are, in general, estimated to be closer to the mean than before. In fact, the highest estimated median salaries are far lower than the actual highest player salaries. This is likely due to players who are in their first few years in MLB performing well, meaning that they have salaries that aren't indicative of their performance. However, when we look at many of the players who have high salaries, we see that the upper quantile estimates of their salaries are more inline with their actual salary. This leads us to believe that we can determine which players are truly overpaid, or underpaid, based on if their salary falls within a $100(1-\alpha)$ % confidence interval found by look at the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantile estimates of their value distribution. We also see that players do have different shapes to their value distribution, even if their median estimates are similar. One such example can be seen in Figure 6.14, where the quantile estimates for Fred Lewis and Steven Drew are shown. Even though they have similar estimates from the .01 to about the .60 quantile levels, the estimates begin to diverge, giving Lewis higher values of quantile estimates in the upper quantiles.

It should be noted that there isn't a large spread among the estimated salaries

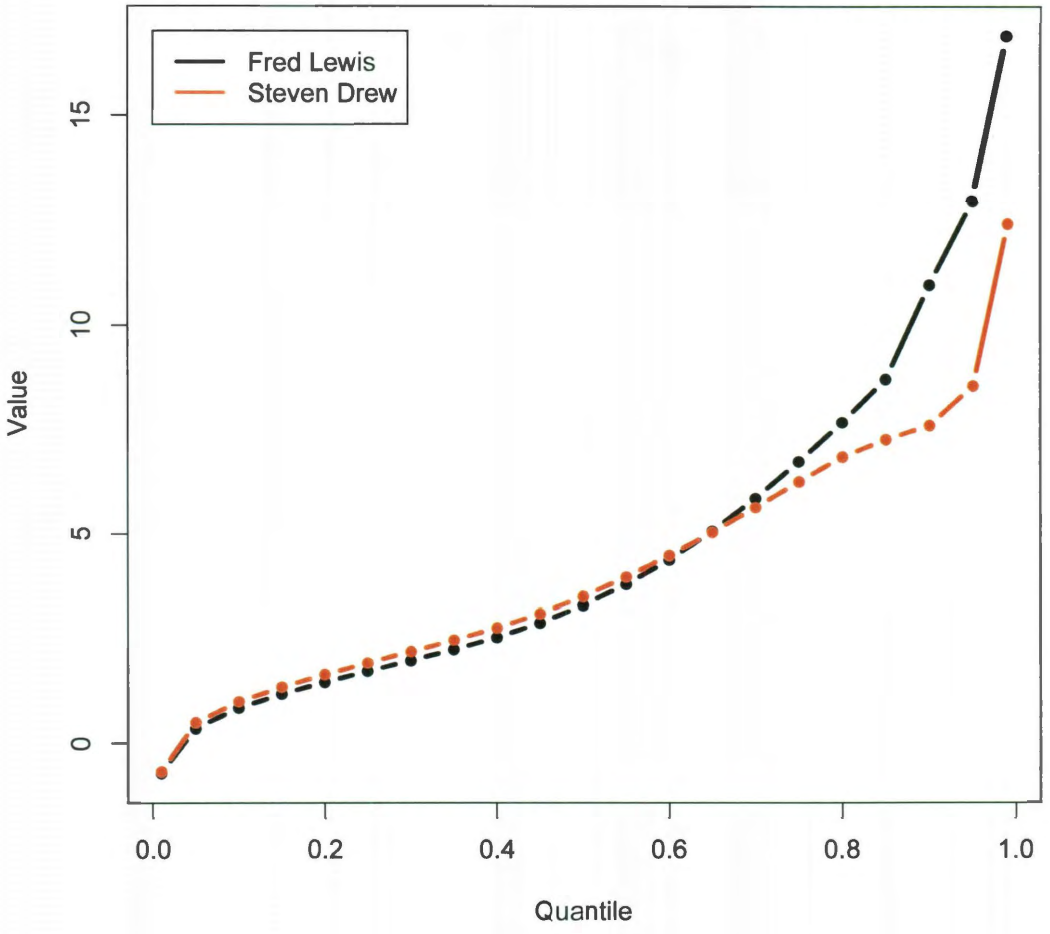


Figure 6.14 : Comparison of the quantile estimates for Fred Lewis and Steven Drew using the AIC-reduced position player model. The estimates for Lewis are shown in black while the estimates for Drew are shown in red.

for either starters or relievers. This is likely due to the high volatility of pitchers. Because of how common it is for elite pitchers becoming below average, and vis versa, salaries for pitchers are commonly non-representative of their abilities. Batters are more stable in their year-to-year performance, which allows their compensation to be more representative of their current performance. Because of this, for the most part, the pitcher models do show a reasonable ordering of player value, but there likely should be more spread in the salaries.

6.4 Discussion

As we have seen in scenarios such as these, where estimating the quantiles of conditional distributions is of interest, there are some benefits to using L_2E quantile regression. For example, in situations where there is not a large amount of contamination, L_2E quantile regression behaves very similarly to KB's quantile regression. However, when there is a contamination present, or if the population is a mixture, L_2E offers robust quantile estimation, giving a better summary of the quantiles of conditional distributions.

Chapter 7

Discussion and Conclusions

Koenker and Bassett's quantile regression is a useful tool to estimate conditional quantiles of data in a wide range of fields, including economics, ecology, sports, and more. However, large amounts of contamination in the data can greatly affect these estimates. We have shown that using Scott's L_2E method to fit a double exponential distribution to data allows for robust quantile estimation. This can be adapted to perform quantile regression, providing robust estimates of conditional quantiles, given that an assumption about the distribution of the residuals is able to be made.

Theoretic results show us how to select the parameters in our L_2E quantile regression criteria to achieve specific quantiles. The theoretic results, particularly the asymptotic analysis, also give us ways of estimating the variance of our estimators and an idea for selecting a constraint on our parameters to achieve unique solutions.

We have also shown semiparametric extensions to the L_2E quantile regression criteria, allowing us to perform non-linear quantile regression. The success of these extensions also creates interest in other extensions to our criteria that can be studied in future research, as well as further study into the two extensions already shown.

Code was written in R, implementing these ideas to perform analysis on both simulated and real data sets. The analysis on simulated data confirmed the assumed behavior of our methods, giving us confidence in our methods. The real data analysis gave us the opportunity to not only compare our method with KB's method, but also to provide applications where our method is useful in providing robust estimates of

conditional distributions.

We believe that quantile regression using L_2E is a method useful in many fields of study, particularly in areas where the population either contains a significant number of outliers or is a mixture density, and warrants further study, consideration, and application.

Bibliography

[1] J. Abrevaya. The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, 26(1):247–257, 2001.

[2] I. Barrodale and F.D.K. Roberts. An improved algorithm for discrete L1 linear approximation. *SIAM Journal on Numerical Analysis*, 10(5):839–848, 1973.

[3] A. Basu, I.R. Harris, N.L. Hjort, and M.C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549, 1998.

[4] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[5] R. Koenker. *Quantile regression*. Cambridge Univ Pr, 2005.

[6] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, 46(1):33–50, 1978.

[7] R. Koenker and K.F. Hallock. Quantile regression. *The Journal of Economic Perspectives*, 15(4):143–156, 2001.

[8] R.W. Koenker and V. D'Orey. Algorithm AS 229: Computing regression quantiles. *Applied Statistics*, 36(3):383–393, 1987.

[9] M.H. Kutner. *Applied linear statistical models*. McGraw-Hill Irwin, 2005.

[10] K. Roger and K. Hallock. Quantile regression. *Journal of Economic Perspectives, pp*, 2001.

[11] E. Ronchetti. Robustness aspects of model choice. *Statistica Sinica*, 7:327–338, 1997.

[12] D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression*, volume 12. Cambridge Univ Pr, 2003.

[13] D.W. Scott. Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43(3):274–285, 2001.

[14] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[15] A.W. Van der Vaart. *Asymptotic statistics*. Number 3. Cambridge Univ Pr, 2000.

[16] K. Yu and M.C. Jones. Local Linear Quantile Regression. *Journal of the American Statistical Association*, 93(441), 1998.

[17] K. Yu, Z. Lu, and J. Stander. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):331–350, 2003.

Appendix A

R Functions

rq.l2e(formula, data, q = .5, c =1, x = FALSE,

y = FALSE, model = FALSE, ...)

Function that fits a linear model using L2E quantile regression. Used similarly to lm(), with q acting as the quantile level and c relating to the parameter in the smooth double exponential distribution. Returns a variable of the l2erq class.

rq.l2efit (x, y, q = .5, c = 1,...)

Function that prepares the data to be used in the L2E quantile regression algorithm.

rq.l2ecal(x,y,c = 1, q = .5)

Function that runs the algorithm to perform L2E quantile regression. lm() is called to as an initial guess of the parameters, normal L2E regression is used to find a robust estimate of the standard deviation of the residuals, then the L2E quantile regression criteria is used to find the coefficient estimates. Finally, the coefficient estimates are rescaled and returned.

rq.l2elog(betas, x,y, a, b, c)

Function that minimizes the L2E quantile regression criteria to find the coefficient estimates.

l2e.nrmlr(betas,x,y)

Function that performs normal L2E regression. Returns coefficient and standard deviation estimates.

print.l2erq (x, ...)

Function that overrides the print values for variables of the l2erq class.

g.log(x, a = 1, b = 1, c = 1)

Function that returns the value of the g-function, that is, the s-curve function using the logistic cdf.

rho.log(x, a = 1, b = 1, c = 1)

Function that returns the value of the rho-function.

fr.log(x, a = 1, b = 1, c = 1)

Function that returns the value of the un-normalized f-function.

f.log(x, a = 1, b = 1, c = 1)

Function that returns the value of the normalized f-function.

fsqr.log(x, a = 1, b = 1, c = 1)

Function that returns the value of the un-normalized squared f-function.

fsq.log(x, a = 1, b = 1, c = 1)

Function that returns the value of the normalized squared f-function.

f.lognrm(x, t,a = 1, b = 1, c = 1,m,s)

Function that returns the value of the smooth double exponential function multiplied by a normal distribution.

theo.log(t, a, b, c, mu, sig)

Function that returns the value of -2 times the integral of f.lognrm().

afnrm.log(a,q,sp,c, mu , sig)

Function that, when used with uniroot(), finds the value of a such that a and b=sp - a give the desired theoretic quantile level.

l2eqr.vartp(x, a = 1, b = 1, c = 1)

Function that returns the numerator in the fraction used for determining the covariance matrix of the coefficient estimates.

l2eqr.varbtm(x, a = 1, b = 1, c = 1)

Function that returns the denominator in the fraction used for

determining the covariance matrix of the coefficient estimates.

l2eqr.var(x,a = 1, b = 1, c = 1)

Function that returns the value of fraction used for determining the covariance matrix of the coefficient estimates.

ffncd1nrm(x,t,a,b,c, m=0, s=1)

Function that calculates the theoretic value of the expectation of the first derivative of the smooth double exponential.

ffncd2nrm(x,t,a,b,c,m=0,s=1)

Function that calculates the theoretic value of the expectation of the second derivative of the smooth double exponential

aicval(t,a,b,c)

Function that calculates the theoretic value of the expectation of the first derivative of the smooth double exponential divided by the expectation of the second derivative of the smooth double exponential. The residuals are assumed to be N(0,1).

Appendix B

Baseball Player Median Value Estimates

The following tables show the median estimates of MLB players from the 2010 season derived from both the full and AIC-reduced models described in Section 6.3. The players actual salary and the difference between the median estimates and the actual salary are also displayed. All values are in millions of dollars.

B.1 Position Players

Player	Team	Salary	Full Model	Difference	AIC	Difference
Stephen Drew	ARI	3.400	2.720	-0.680	3.502	0.102
Kelly Johnson	ARI	2.350	4.901	2.551	6.969	4.619
Adam LaRoche	ARI	4.500	3.882	-0.618	5.602	1.102
Mark Reynolds	ARI	0.833	7.045	6.211	8.973	8.140
Chris Snyder	ARI	5.250	6.122	0.872	5.223	-0.027
Justin Upton	ARI	0.708	4.778	4.070	5.997	5.289
Chris Young	ARI	3.450	7.176	3.726	8.797	5.347
Melky Cabrera	ATL	3.100	4.140	1.040	2.884	-0.216
Yunel Escobar	ATL	0.435	4.998	4.563	4.982	4.547
Troy Glaus	ATL	1.750	6.338	4.588	6.805	5.055
Omar Infante	ATL	2.225	3.181	0.956	3.524	1.299
Chipper Jones	ATL	14.000	6.604	-7.396	6.236	-7.764
Brian McCann	ATL	5.700	8.211	2.511	7.612	1.912
Martin Prado	ATL	0.440	4.179	3.739	5.575	5.135
Cesar Izturis	BAL	2.600	2.536	-0.064	1.839	-0.761
Adam Jones	BAL	0.465	2.302	1.837	2.713	2.248
Nick Markakis	BAL	7.100	6.496	-0.604	6.691	-0.409
Luke Scott	BAL	4.050	6.069	2.019	6.824	2.774
Miguel Tejada	BAL	6.000	5.906	-0.094	5.138	-0.862
Ty Wigginton	BAL	3.500	5.459	1.959	5.405	1.905

Player	Team	Salary	Full Model	Difference	AIC	Difference
Adrian Beltre	BOS	9.000	6.457	-2.543	4.956	-4.044
J.D. Drew	BOS	14.000	6.107	-7.893	6.717	-7.283
Bill Hall	BOS	8.525	2.672	-5.853	3.753	-4.772
Victor Martinez	BOS	7.700	6.146	-1.554	5.063	-2.637
David Ortiz	BOS	13.000	9.063	-3.937	8.986	-4.014
Dustin Pedroia	BOS	3.750	3.778	0.028	4.045	0.295
Marco Scutaro	BOS	5.500	5.667	0.167	6.663	1.163
Kevin Youkilis	BOS	9.375	5.123	-4.252	5.969	-3.406
Gordon Beckham	CHA	0.445	1.858	1.413	2.802	2.357
Paul Konerko	CHA	12.000	8.715	-3.285	8.818	-3.182
Mark Kotsay	CHA	1.500	3.534	2.034	2.519	1.019
Jayson Nix	CHA	0.420	1.758	1.338	1.991	1.571
Juan Pierre	CHA	7.000	4.301	-2.699	3.076	-3.924
A.J. Pierzynski	CHA	6.750	2.489	-4.261	0.901	-5.849
Carlos Quentin	CHA	3.200	6.644	3.444	5.875	2.675
Alexei Ramirez	CHA	1.225	3.518	2.293	3.361	2.136
Alexis Rios	CHA	10.200	6.002	-4.198	4.877	-5.323
Omar Vizquel	CHA	1.375	2.047	0.672	2.049	0.674
Marlon Byrd	CHN	3.000	3.351	0.351	4.705	1.705
Kosuke Fukudome	CHN	14.000	4.047	-9.953	4.509	-9.491
Derrek Lee	CHN	13.250	6.548	-6.702	8.181	-5.069
Xavier Nady	CHN	3.300	1.505	-1.795	2.494	-0.806
Aramis Ramirez	CHN	16.750	5.527	-11.223	5.039	-11.711
Alfonso Soriano	CHN	19.000	3.661	-15.339	4.191	-14.809
Geovany Soto	CHN	0.575	5.493	4.918	6.333	5.758
Ryan Theriot	CHN	2.600	4.094	1.494	4.221	1.621
Jay Bruce	CIN	0.440	5.143	4.703	6.270	5.830
Orlando Cabrera	CIN	2.020	2.735	0.715	3.205	1.185
Jonny Gomes	CIN	0.800	4.001	3.201	5.241	4.441
Ramon Hernandez	CIN	3.868	2.285	-1.584	2.174	-1.694
Brandon Phillips	CIN	6.938	4.237	-2.700	4.348	-2.590
Scott Rolen	CIN	7.667	4.740	-2.926	4.802	-2.865
Joey Votto	CIN	0.525	9.691	9.166	10.202	9.677
Russell Branyan	CLE	1.500	3.980	2.480	5.117	3.617
Asdrubal Cabrera	CLE	0.445	0.605	0.160	0.827	0.382
Shin-Soo Choo	CLE	0.461	8.400	7.939	8.010	7.549
Travis Hafner	CLE	11.500	4.526	-6.974	4.727	-6.773
Austin Kearns	CLE	0.750	3.803	3.053	5.202	4.452
Jhonny Peralta	CLE	4.850	4.900	0.050	5.436	0.586
Clint Barmes	COL	3.325	4.743	1.418	3.610	0.285

Player	Team	Salary	Full Model	Difference	AIC	Difference
Todd Helton	COL	17.775	5.160	-12.615	6.525	-11.250
Melvin Mora	COL	1.275	2.546	1.271	2.129	0.854
Miguel Olivo	COL	2.000	1.741	-0.259	2.183	0.183
Ryan Spilborghs	COL	1.300	1.936	0.636	2.924	1.624
Troy Tulowitzki	COL	3.500	6.594	3.094	6.229	2.729
Miguel Cabrera	DET	20.000	14.072	-5.928	10.447	-9.553
Johnny Damon	DET	8.000	4.695	-3.305	6.179	-1.821
Brandon Inge	DET	6.600	2.927	-3.673	3.471	-3.129
Magglio Ordonez	DET	17.826	5.363	-12.463	5.384	-12.442
Ryan Raburn	DET	0.438	2.379	1.941	3.418	2.980
Ramon Santiago	DET	1.250	1.554	0.304	2.176	0.926
Jorge Cantu	FLO	6.000	2.593	-3.407	3.253	-2.747
Chris Coghlan	FLO	0.475	1.887	1.412	3.518	3.043
Ronny Paulino	FLO	1.100	3.291	2.191	2.938	1.838
Hanley Ramirez	FLO	7.000	8.566	1.566	7.504	0.504
Cody Ross	FLO	4.450	3.745	-0.705	4.689	0.239
Dan Uggla	FLO	7.800	8.059	0.259	10.276	2.476
Lance Berkman	HOU	14.500	7.367	-7.133	6.768	-7.732
Michael Bourn	HOU	2.400	3.812	1.412	4.365	1.965
Pedro Feliz	HOU	4.500	3.152	-1.348	2.110	-2.390
Jeff Keppinger	HOU	1.150	4.796	3.646	4.512	3.362
Carlos Lee	HOU	19.000	6.717	-12.283	4.932	-14.068
Hunter Pence	HOU	3.500	5.792	2.292	5.905	2.405
Mike Aviles	KCA	0.429	3.226	2.797	3.538	3.109
Yuniesky Betancourt	KCA	3.300	3.558	0.258	2.521	-0.779
Billy Butler	KCA	0.470	7.666	7.196	7.329	6.859
Alberto Callaspo	KCA	0.460	5.196	4.736	3.787	3.327
David DeJesus	KCA	4.700	2.289	-2.411	2.455	-2.245
Jose Guillen	KCA	12.000	3.755	-8.245	3.916	-8.084
Jason Kendall	KCA	2.250	3.048	0.798	2.464	0.214
Scott Podsednik	KCA	1.650	1.439	-0.211	1.070	-0.580
Bobby Abreu	LAA	9.000	7.106	-1.894	8.407	-0.593
Erick Aybar	LAA	2.050	1.833	-0.217	2.115	0.065
Torii Hunter	LAA	18.500	6.777	-11.723	6.203	-12.297
Howie Kendrick	LAA	1.750	2.083	0.333	1.855	0.105
Hideki Matsui	LAA	6.000	6.790	0.790	6.592	0.592
Mike Napoli	LAA	3.600	4.187	0.587	5.002	1.402
Juan Rivera	LAA	4.250	5.223	0.973	4.606	0.356
Casey Blake	LAN	6.250	2.995	-3.255	4.231	-2.019
Jamey Carroll	LAN	1.536	3.376	1.840	4.382	2.846

Player	Team	Salary	Full Model	Difference	AIC	Difference
Andre Ethier	LAN	5.750	6.986	1.236	6.596	0.846
Rafael Furcal	LAN	9.500	3.322	-6.178	3.137	-6.363
Matt Kemp	LAN	4.000	4.107	0.107	4.667	0.667
James Loney	LAN	3.100	5.690	2.590	4.733	1.633
Russell Martin	LAN	5.050	5.241	0.191	5.271	0.221
Ryan Braun	MIL	1.288	6.659	5.371	7.526	6.238
Prince Fielder	MIL	11.000	12.637	1.637	12.306	1.306
Corey Hart	MIL	4.800	4.673	-0.127	5.570	0.770
Casey McGehee	MIL	0.427	6.109	5.682	5.664	5.237
Rickie Weeks	MIL	2.750	5.602	2.852	8.966	6.216
Michael Cuddyer	MIN	9.417	7.011	-2.406	6.375	-3.042
J.J. Hardy	MIN	5.100	2.117	-2.983	2.399	-2.701
Orlando Hudson	MIN	5.000	3.406	-1.594	4.787	-0.213
Jason Kubel	MIN	4.100	6.358	2.258	6.152	2.052
Joe Mauer	MIN	12.500	8.184	-4.316	7.323	-5.177
Denard Span	MIN	0.750	3.474	2.724	3.234	2.484
Delmon Young	MIN	2.600	4.736	2.136	3.896	1.296
Robinson Cano	NYA	9.000	9.215	0.215	7.385	-1.615
Brett Gardner	NYA	0.453	5.371	4.918	6.951	6.498
Curtis Granderson	NYA	5.500	4.298	-1.202	4.805	-0.695
Derek Jeter	NYA	22.600	7.649	-14.951	8.506	-14.094
Jorge Posada	NYA	13.100	4.880	-8.220	5.595	-7.505
Alex Rodriguez	NYA	33.000	6.613	-26.387	6.554	-26.446
Nick Swisher	NYA	6.850	4.802	-2.048	6.580	-0.270
Mark Teixeira	NYA	20.625	10.896	-9.729	11.651	-8.974
Rod Barajas	NYN	0.500	3.720	3.220	2.582	2.082
Jason Bay	NYN	8.625	2.594	-6.031	3.284	-5.341
Jeff Francoeur	NYN	5.000	4.628	-0.372	3.674	-1.326
Angel Pagan	NYN	1.500	3.255	1.755	2.736	1.236
Jose Reyes	NYN	9.375	2.633	-6.742	1.534	-7.841
David Wright	NYN	10.250	6.140	-4.110	6.735	-3.515
Rajai Davis	OAK	1.350	2.775	1.425	2.698	1.348
Mark Ellis	OAK	5.500	3.458	-2.042	3.324	-2.176
Kevin Kouzmanoff	OAK	3.100	3.834	0.734	3.464	0.364
Kurt Suzuki	OAK	0.420	6.009	5.589	4.083	3.663
Ryan Sweeney	OAK	0.420	2.537	2.117	2.551	2.131
Ryan Howard	PHI	19.000	7.108	-11.892	6.930	-12.070
Raul Ibanez	PHI	12.167	6.419	-5.748	5.885	-6.282
Placido Polanco	PHI	5.167	4.521	-0.646	5.012	-0.155
Jimmy Rollins	PHI	8.500	4.868	-3.632	4.173	-4.327

Player	Team	Salary	Full Model	Difference	AIC	Difference
Carlos Ruiz	PHI	1.900	5.954	4.054	4.652	2.752
Chase Utley	PHI	15.286	6.864	-8.421	7.199	-8.087
Shane Victorino	PHI	5.000	5.594	0.594	4.206	-0.794
Jayson Werth	PHI	7.500	6.876	-0.624	9.198	1.698
Ronny Cedeno	PIT	1.125	0.320	-0.805	0.410	-0.715
Ryan Doumit	PIT	3.650	4.283	0.633	4.050	0.400
Garrett Jones	PIT	0.425	5.291	4.866	5.535	5.110
Andrew McCutchen	PIT	0.422	5.383	4.961	6.558	6.136
Lastings Milledge	PIT	0.452	1.885	1.433	1.832	1.380
David Eckstein	SDN	1.000	1.599	0.599	1.537	0.537
Adrian Gonzalez	SDN	4.875	13.859	8.984	9.979	5.104
Jerry Hairston	SDN	2.125	3.696	1.571	3.165	1.040
Chase Headley	SDN	0.428	4.266	3.838	5.898	5.471
Yorvit Torrealba	SDN	0.750	3.086	2.336	3.034	2.284
Chone Figgins	SEA	8.500	2.546	-5.954	3.042	-5.458
Franklin Gutierrez	SEA	2.312	4.121	1.808	4.664	2.351
Casey Kotchman	SEA	3.518	4.995	1.478	3.525	0.008
Jose Lopez	SEA	3.000	4.244	1.244	3.172	0.172
Ichiro Suzuki	SEA	18.000	5.144	-12.856	4.203	-13.797
Aubrey Huff	SFN	3.000	8.474	5.474	8.775	5.775
Fred Lewis	TOR	0.455	1.744	1.289	3.295	2.840
Bengie Molina	SFN	4.500	3.127	-1.373	1.636	-2.864
Aaron Rowand	SFN	13.600	2.570	-11.030	2.229	-11.371
Freddy Sanchez	SFN	6.000	2.178	-3.822	2.880	-3.120
Pablo Sandoval	SFN	0.465	6.336	5.871	4.399	3.934
Andres Torres	SFN	0.426	2.125	1.699	3.133	2.707
Juan Uribe	SFN	3.250	6.652	3.402	5.300	2.050
Matt Holliday	SLN	16.333	8.416	-7.917	7.768	-8.565
Felipe Lopez	SLN	1.000	3.735	2.735	4.576	3.576
Ryan Ludwick	SLN	5.450	3.707	-1.743	5.011	-0.439
Yadier Molina	SLN	4.312	5.313	1.000	3.367	-0.946
Albert Pujols	SLN	14.596	17.571	2.975	11.888	-2.708
Colby Rasmus	SLN	0.418	5.174	4.756	6.566	6.148
Brendan Ryan	SLN	0.425	2.572	2.147	1.741	1.316
Skip Schumaker	SLN	2.000	4.732	2.732	5.507	3.507
Jason Bartlett	TBA	4.000	1.999	-2.001	2.909	-1.091
Pat Burrell	TBA	9.000	5.172	-3.828	5.988	-3.012
Carl Crawford	TBA	10.000	3.546	-6.454	3.546	-6.454
Evan Longoria	TBA	0.950	7.010	6.060	6.797	5.847
Carlos Pena	TBA	10.125	7.738	-2.387	9.067	-1.058

Player	Team	Salary	Full Model	Difference	AIC	Difference
B.J. Upton	TBA	3.000	4.388	1.388	6.298	3.298
Ben Zobrist	TBA	0.438	6.516	6.078	7.706	7.268
Elvis Andrus	TEX	0.418	2.971	2.552	4.154	3.735
Julio Borbon	TEX	0.600	1.526	0.926	1.582	0.982
Nelson Cruz	TEX	0.440	4.118	3.678	3.556	3.116
Vladimir Guerrero	TEX	5.500	7.695	2.195	5.722	0.222
Josh Hamilton	TEX	3.250	5.443	2.193	5.930	2.680
Ian Kinsler	TEX	4.200	5.831	1.631	6.502	2.302
David Murphy	TEX	0.428	4.342	3.914	4.580	4.152
Michael Young	TEX	13.175	6.308	-6.867	7.075	-6.100
Jose Bautista	TOR	2.400	11.026	8.626	10.944	8.544
John Buck	TOR	2.000	2.033	0.033	3.324	1.324
Edwin Encarnacion	TOR	5.175	4.548	-0.627	4.371	-0.804
Alex Gonzalez	TOR	2.750	3.135	0.385	3.097	0.347
Aaron Hill	TOR	4.000	6.497	2.497	5.917	1.917
Adam Lind	TOR	0.550	3.128	2.578	3.776	3.226
Lyle Overbay	TOR	7.950	5.944	-2.006	6.942	-1.008
Vernon Wells	TOR	15.688	6.286	-9.401	5.041	-10.647
Adam Dunn	WAS	12.000	6.771	-5.229	8.142	-3.858
Cristian Guzman	WAS	8.000	1.716	-6.284	1.872	-6.128
Adam Kennedy	WAS	1.250	3.930	2.680	3.906	2.656
Nyjer Morgan	WAS	0.426	-0.661	-1.088	-0.594	-1.020
Ivan Rodriguez	WAS	3.000	2.361	-0.639	1.461	-1.539
Josh Willingham	WAS	4.600	5.807	1.207	6.431	1.831
Ryan Zimmerman	WAS	6.350	8.049	1.699	8.561	2.211

B.2 Pitchers

B.2.1 Starting Pitchers

Player	Team	Salary	Full Model	Difference	AIC	Difference
Danny Haren	ARI	8.250	4.678	-3.572	5.952	-2.298
Edwin Jackson	ARI	4.600	6.183	1.583	4.749	0.149
Rodrigo Lopez	ARI	0.650	3.171	2.521	5.065	4.415
Tommy Hanson	ATL	0.435	7.058	6.623	6.284	5.849
Tim Hudson	ATL	9.000	7.684	-1.316	5.801	-3.199
Jair Jurrjens	ATL	0.480	4.353	3.873	5.098	4.618
Kenshin Kawakami	ATL	7.334	3.187	-4.146	4.104	-3.230

Player	Team	Salary	Full Model	Difference	AIC	Difference
Derek Lowe	ATL	15.000	6.272	-8.728	4.977	-10.023
Jeremy Guthrie	BAL	3.000	5.222	2.222	6.468	3.468
Brian Matusz	BAL	1.300	5.983	4.683	5.045	3.745
Kevin Millwood	BAL	12.000	3.131	-8.869	4.514	-7.486
Josh Beckett	BOS	12.100	3.813	-8.287	5.337	-6.763
Clay Buchholz	BOS	0.443	7.570	7.127	5.477	5.034
John Lackey	BOS	18.700	7.438	-11.262	5.212	-13.488
Jon Lester	BOS	3.750	8.608	4.858	5.409	1.659
Daisuke Matsuzaka	BOS	8.333	6.452	-1.881	4.708	-3.626
Tim Wakefield	BOS	3.500	4.053	0.553	5.903	2.403
Mark Buehrle	CHA	14.000	4.406	-9.594	5.594	-8.406
John Danks	CHA	3.450	6.561	3.111	5.483	2.033
Gavin Floyd	CHA	2.750	6.004	3.254	5.188	2.438
Freddy Garcia	CHA	1.000	4.901	3.901	5.761	4.761
Jake Peavy	CHA	15.000	2.964	-12.036	6.089	-8.911
Ryan Dempster	CHN	13.500	6.984	-6.516	4.825	-8.675
Tom Gorzelanny	CHN	0.800	-1.150	-1.950	3.685	2.885
Ted Lilly	CHN	13.000	2.687	-10.313	6.413	-6.587
Carlos Silva	CHN	12.750	5.188	-7.562	6.900	-5.850
Randy Wells	CHN	0.427	5.724	5.297	4.914	4.487
Carlos Zambrano	CHN	18.875	8.710	-10.165	4.157	-14.718
Bronson Arroyo	CIN	11.625	5.716	-5.909	6.100	-5.525
Homer Bailey	CIN	0.418	3.424	3.006	5.347	4.929
Johnny Cueto	CIN	0.445	5.471	5.026	5.996	5.551
Aaron Harang	CIN	12.500	3.966	-8.534	5.181	-7.319
Fausto Carmona	CLE	5.088	7.092	2.005	5.191	0.103
Justin Masterson	CLE	0.427	6.981	6.554	4.438	4.011
Jake Westbrook	CLE	11.000	6.298	-4.702	5.290	-5.710
Aaron Cook	COL	9.625	4.753	-4.872	4.442	-5.183
Jorge de la Rosa	COL	5.600	3.845	-1.755	4.725	-0.875
Jeff Francis	COL	5.750	3.638	-2.112	6.304	0.554
Jason Hammel	COL	1.900	4.803	2.903	5.905	4.005
Ubaldo Jimenez	COL	1.250	8.833	7.583	5.223	3.973
Jeremy Bonderman	DET	12.500	4.965	-7.535	5.278	-7.222
Rick Porcello	DET	1.920	3.530	1.610	5.940	4.020
Max Scherzer	DET	1.500	6.238	4.738	5.351	3.851
Justin Verlander	DET	6.850	7.369	0.519	5.944	-0.906
Josh Johnson	FLO	3.750	7.025	3.275	6.633	2.883
Ricky Nolasco	FLO	3.800	4.361	0.561	6.381	2.581
Nate Robertson	FLO	10.000	4.007	-5.993	4.746	-5.254

Player	Team	Salary	Full Model	Difference	AIC	Difference
Anibal Sanchez	FLO	1.250	7.834	6.584	4.954	3.704
Chris Volstad	FLO	0.420	6.121	5.701	5.246	4.826
Brian Moehler	HOU	3.000	3.422	0.422	4.226	1.226
Brett Myers	HOU	3.100	6.810	3.710	5.846	2.746
Roy Oswalt	HOU	15.000	5.790	-9.210	6.235	-8.765
Felipe Paulino	HOU	0.415	4.730	4.315	3.643	3.228
Wandy Rodriguez	HOU	5.000	6.760	1.760	5.277	0.277
Brian Bannister	KCA	2.300	3.254	0.954	4.240	1.940
Kyle Davies	KCA	1.800	5.286	3.486	3.920	2.120
Zack Greinke	KCA	7.250	6.157	-1.093	5.898	-1.352
Luke Hochevar	KCA	1.760	3.965	2.205	5.394	3.634
Gil Meche	KCA	12.400	4.064	-8.336	2.828	-9.572
Scott Kazmir	LAA	8.000	5.049	-2.951	3.428	-4.572
Joel Pineiro	LAA	8.000	4.765	-3.235	6.274	-1.726
Ervin Santana	LAA	6.000	6.657	0.657	5.622	-0.378
Joe Saunders	LAA	3.700	5.199	1.499	4.710	1.010
Jered Weaver	LAA	4.265	5.347	1.082	6.244	1.979
Chad Billingsley	LAN	3.850	7.963	4.113	5.213	1.363
Clayton Kershaw	LAN	0.440	6.652	6.212	5.036	4.596
Hiroki Kuroda	LAN	15.433	6.194	-9.240	5.775	-9.659
Vicente Padilla	LAN	5.025	1.973	-3.052	7.003	1.978
David Bush	MIL	4.215	3.942	-0.273	4.313	0.098
Yovani Gallardo	MIL	0.450	7.237	6.787	4.813	4.363
Manny Parra	MIL	0.440	5.916	5.476	3.335	2.895
Jeff Suppan	MIL	12.750	5.112	-7.638	4.615	-8.135
Randy Wolf	MIL	8.800	6.270	-2.530	4.569	-4.231
Scott Baker	MIN	3.000	5.152	2.152	6.136	3.136
Nick Blackburn	MIN	0.750	3.662	2.912	5.612	4.862
Francisco Liriano	MIN	1.600	7.511	5.911	5.959	4.359
Carl Pavano	MIN	7.000	6.375	-0.625	6.784	-0.216
Kevin Slowey	MIN	0.470	5.281	4.811	6.900	6.430
A.J. Burnett	NYA	16.500	6.348	-10.152	4.537	-11.963
Philip Hughes	NYA	0.447	5.518	5.071	5.547	5.100
Andy Pettitte	NYA	11.750	5.271	-6.479	6.036	-5.714
C.C. Sabathia	NYA	24.286	8.035	-16.250	5.959	-18.327
Javier Vazquez	NYA	11.500	4.174	-7.326	4.692	-6.808
Mike Pelfrey	NYN	0.500	0.111	-0.389	5.276	4.776
Oliver Perez	NYN	12.000	4.406	-7.594	0.194	-11.806
Johan Santana	NYN	20.145	5.017	-15.128	6.031	-14.113
Dallas Braden	OAK	0.420	5.145	4.725	6.245	5.825

Player	Team	Salary	Full Model	Difference	AIC	Difference
Ben Sheets	OAK	10.000	2.559	-7.441	4.891	-5.109
Joe Blanton	PHI	3.000	4.305	1.305	5.962	2.962
Roy Halladay	PHI	15.750	6.782	-8.968	7.509	-8.241
Cole Hamels	PHI	6.650	5.689	-0.961	5.876	-0.774
J.A. Happ	PHI	0.470	4.689	4.219	4.103	3.633
Kyle Kendrick	PHI	0.480	4.120	3.640	5.601	5.121
Jamie Moyer	PHI	8.000	2.340	-5.660	7.302	-0.698
Zach Duke	PIT	4.300	2.447	-1.853	4.408	0.108
Paul Maholm	PIT	5.000	6.642	1.642	4.656	-0.344
Charlie Morton	PIT	0.422	1.057	0.634	4.789	4.367
Ross Ohlendorf	PIT	0.439	3.732	3.293	4.793	4.354
Kevin Correia	SDN	3.600	5.427	1.827	4.137	0.537
Jon Garland	SDN	4.700	7.053	2.353	4.292	-0.408
Clayton Richard	SDN	0.424	6.631	6.207	4.720	4.296
Felix Hernandez	SEA	7.200	6.664	-0.536	6.184	-1.016
Cliff Lee	SEA	9.000	4.943	-4.057	7.762	-1.238
Ryan Rowland-Smith	SEA	0.440	2.279	1.839	4.390	3.950
Matt Cain	SFN	4.583	6.113	1.529	6.094	1.511
Tim Lincecum	SFN	9.000	7.504	-1.496	5.130	-3.870
Jonathan Sanchez	SFN	2.100	6.671	4.571	4.497	2.397
Barry Zito	SFN	18.500	6.318	-12.182	4.279	-14.221
Chris Carpenter	SLN	15.841	7.603	-8.238	6.381	-9.460
Kyle Lohse	SLN	9.188	4.561	-4.626	4.280	-4.907
Adam Wainwright	SLN	4.838	7.939	3.101	6.421	1.584
Matt Garza	TBA	3.350	-2.533	-5.883	5.761	2.411
Jeff Niemann	TBA	1.032	5.403	4.371	5.494	4.462
David Price	TBA	1.835	7.045	5.210	5.606	3.771
James Shields	TBA	2.500	3.975	1.475	5.315	2.815
Scott Feldman	TEX	2.425	5.235	2.810	4.915	2.490
Rich Harden	TEX	6.500	4.059	-2.441	3.169	-3.331
Colby Lewis	TEX	1.750	5.971	4.221	5.551	3.801
C.J. Wilson	TEX	3.100	8.287	5.187	4.990	1.890
Shaun Marcum	TOR	0.850	5.510	4.660	6.715	5.865
John Lannan	WAS	0.458	5.486	5.028	4.947	4.489

B.2.2 Relief Pitchers

Player	Team	Salary	Full Model	Difference	AIC	Difference
Blaine Boyer	ARI	0.725	1.006	0.281	1.796	1.071
Aaron Heilman	ARI	2.150	1.686	-0.464	1.923	-0.227
Bobby Howry	ARI	2.000	1.406	-0.594	1.850	-0.150
Chad Qualls	ARI	4.185	2.294	-1.891	1.872	-2.313
Jesse Chavez	ATL	0.415	1.301	0.886	1.863	1.448
Peter Moylan	ATL	1.150	1.225	0.075	1.598	0.448
Eric O'Flaherty	ATL	0.440	1.185	0.745	2.187	1.747
Takashi Saito	ATL	3.200	1.407	-1.793	2.443	-0.757
Billy Wagner	ATL	6.750	3.434	-3.316	2.693	-4.057
Matt Albers	BAL	0.680	1.014	0.334	1.874	1.194
Mike Gonzalez	BAL	6.000	1.081	-4.919	1.559	-4.441
Mark Hendrickson	BAL	1.200	1.170	-0.030	2.120	0.920
Jim Johnson	BAL	0.440	1.283	0.843	2.718	2.278
Cla Meredith	BAL	0.850	1.462	0.612	2.384	1.534
Will Ohman	BAL	1.350	1.171	-0.179	1.568	0.218
Koji Uehara	BAL	5.000	2.264	-2.736	3.108	-1.892
Scott Atchison	BOS	0.420	1.058	0.638	2.322	1.902
Daniel Bard	BOS	0.415	1.141	0.726	2.365	1.949
Manny Delcarmen	BOS	0.905	1.167	0.262	1.412	0.507
Hideki Okajima	BOS	2.750	1.363	-1.387	1.539	-1.211
Jonathan Papelbon	BOS	9.350	3.474	-5.876	1.838	-7.512
Ramon Ramirez	BOS	1.155	1.179	0.024	2.193	1.038
Bobby Jenks	CHA	7.500	2.924	-4.576	2.274	-5.226
Scott Linebrink	CHA	5.000	1.231	-3.769	2.486	-2.514
J.J. Putz	CHA	3.000	1.741	-1.259	2.471	-0.529
Matt Thornton	CHA	2.250	1.824	-0.426	2.412	0.162
Randy Williams	CHA	0.415	0.639	0.224	0.421	0.006
John Grabow	CHN	2.700	1.420	-1.280	1.563	-1.137
Carlos Marmol	CHN	2.125	3.080	0.955	1.706	-0.419
Sean Marshall	CHN	0.950	1.533	0.583	2.271	1.321
Francisco Cordero	CIN	12.125	3.550	-8.575	1.727	-10.398
Mike Lincoln	CIN	2.500	1.164	-1.336	2.017	-0.483
Nick Masset	CIN	1.035	1.358	0.323	1.986	0.951
Micah Owings	CIN	0.440	0.891	0.451	1.344	0.904
Arthur Rhodes	CIN	2.000	1.462	-0.538	2.458	0.458
Aaron Laffey	CLE	0.422	0.721	0.299	1.757	1.335
Jensen Lewis	CLE	0.422	1.104	0.682	1.879	1.457
Chris Perez	CLE	0.424	2.194	1.770	2.303	1.879
Rafael Perez	CLE	0.795	1.153	0.358	1.948	1.153

Player	Team	Salary	Full Model	Difference	AIC	Difference
Joe Smith	CLE	0.428	1.130	0.702	1.651	1.223
Kerry Wood	CLE	10.500	1.361	-9.139	1.443	-9.057
Jamey Wright	CLE	0.900	0.884	-0.016	2.120	1.220
Matt Belisle	COL	0.850	1.397	0.547	2.675	1.825
Rafael Betancourt	COL	3.775	1.714	-2.061	3.091	-0.684
Manuel Corpas	COL	2.750	1.833	-0.917	2.011	-0.739
Randy Flores	COL	0.650	1.312	0.662	1.902	1.252
Huston Street	COL	7.200	2.506	-4.694	2.621	-4.579
Phil Coke	DET	0.425	1.514	1.089	1.977	1.552
Brad Thomas	DET	1.000	0.828	-0.172	2.062	1.062
Jose Valverde	DET	6.886	2.498	-4.388	1.966	-4.920
Joel Zumaya	DET	0.915	1.192	0.277	2.668	1.753
Clay Hensley	FLO	0.425	1.437	1.012	2.266	1.841
Leo Nunez	FLO	2.000	3.120	1.120	2.297	0.297
Renyel Pinto	FLO	1.075	0.757	-0.318	2.285	1.210
Jose Veras	FLO	0.550	1.109	0.559	1.685	1.135
Tim Byrdak	HOU	1.600	1.353	-0.247	1.707	0.107
Jeff Fulchino	HOU	0.425	1.228	0.803	1.973	1.548
Matt Lindstrom	HOU	1.625	2.622	0.997	1.801	0.176
Brandon Lyon	HOU	4.250	2.292	-1.958	1.877	-2.373
Chris Sampson	HOU	0.815	1.436	0.621	2.411	1.596
Kyle Farnsworth	KCA	4.500	1.187	-3.313	2.673	-1.827
Joakim Soria	KCA	3.000	3.625	0.625	2.762	-0.238
Robinson Tejeda	KCA	0.950	1.178	0.228	1.899	0.949
Jason Bulger	LAA	0.418	0.996	0.578	1.671	1.253
Brian Fuentes	LAA	9.000	2.461	-6.539	2.411	-6.589
Kevin Jepsen	LAA	0.415	1.133	0.718	1.766	1.351
Fernando Rodney	LAA	5.500	1.921	-3.579	1.827	-3.673
Scot Shields	LAA	5.350	0.766	-4.584	1.071	-4.279
Brian Stokes	LAA	0.435	0.787	0.352	0.127	-0.308
Jonathan Broxton	LAN	4.000	2.656	-1.344	1.679	-2.321
Hong-Chih Kuo	LAN	0.950	1.818	0.868	2.823	1.873
Ramon Ortiz	LAN	1.000	1.006	0.006	1.564	0.564
George Sherrill	LAN	4.500	1.366	-3.134	1.082	-3.418
Ramon Troncoso	LAN	0.416	1.155	0.739	2.241	1.825
Jeff Weaver	LAN	0.800	1.222	0.422	1.913	1.113
Todd Coffey	MIL	2.025	1.294	-0.731	2.061	0.036
LaTroy Hawkins	MIL	3.250	1.546	-1.704	2.178	-1.072
Trevor Hoffman	MIL	7.500	2.028	-5.472	1.709	-5.791
David Riske	MIL	4.500	1.176	-3.324	2.577	-1.923

Player	Team	Salary	Full Model	Difference	AIC	Difference
Claudio Vargas	MIL	0.900	1.269	0.369	1.648	0.748
Carlos Villanueva	MIL	0.950	1.191	0.241	2.376	1.426
Jesse Crain	MIN	2.000	1.096	-0.904	2.300	0.300
Brian Duensing	MIN	0.417	0.269	-0.148	2.496	2.078
Matt Guerrier	MIN	3.150	1.461	-1.689	2.291	-0.859
Jose Mijares	MIN	0.430	1.410	0.980	2.538	2.108
Jon Rauch	MIN	2.900	2.508	-0.392	2.678	-0.222
Joba Chamberlain	NYA	0.488	1.584	1.096	2.310	1.822
Damaso Marte	NYA	4.000	0.857	-3.143	1.891	-2.109
Sergio Mitre	NYA	0.850	0.890	0.040	2.573	1.723
Chan Ho Park	NYA	1.200	1.207	0.007	2.430	1.230
Mariano Rivera	NYA	15.000	3.103	-11.897	3.083	-11.917
David Robertson	NYA	0.427	1.187	0.761	1.502	1.076
Pedro Feliciano	NYN	2.900	1.449	-1.451	1.679	-1.221
Ryota Igarashi	NYN	1.250	1.170	-0.080	1.625	0.375
Francisco Rodriguez	NYN	12.167	2.518	-9.649	2.347	-9.820
Hisanori Takahashi	NYN	1.000	1.022	0.022	1.972	0.972
Andrew Bailey	OAK	0.435	2.599	2.164	2.640	2.205
Craig Breslow	OAK	0.425	1.542	1.117	2.140	1.715
Chad Gaudin	OAK	0.700	1.035	0.335	2.161	1.461
Michael Wuertz	OAK	2.200	1.432	-0.768	1.556	-0.644
Danys Baez	PHI	2.500	1.267	-1.233	1.715	-0.785
Jose Contreras	PHI	1.500	1.805	0.305	2.499	0.999
Chad Durbin	PHI	2.125	1.130	-0.995	2.364	0.239
Brad Lidge	PHI	12.000	2.498	-9.502	1.905	-10.095
Ryan Madson	PHI	4.833	1.673	-3.160	2.816	-2.017
J.C. Romero	PHI	4.250	0.910	-3.340	1.170	-3.080
D.J. Carrasco	PIT	0.950	0.845	-0.105	2.198	1.248
Brendan Donnelly	PIT	1.350	0.988	-0.362	0.985	-0.365
Octavio Dotel	PIT	3.250	2.390	-0.860	1.788	-1.462
Joel Hanrahan	PIT	0.453	1.649	1.196	2.500	2.047
Javier Lopez	PIT	0.775	1.283	0.508	2.384	1.609
Jack Taschner	PIT	0.835	1.161	0.326	1.649	0.814
Mike Adams	SDN	1.000	1.211	0.211	2.507	1.507
Heath Bell	SDN	4.000	3.809	-0.191	2.299	-1.701
Luke Gregerson	SDN	0.416	1.710	1.293	2.612	2.195
Edward Mujica	SDN	0.420	1.405	0.986	3.222	2.803
Tim Stauffer	SDN	0.415	0.622	0.207	2.389	1.974
David Aardsma	SEA	2.750	2.862	0.112	1.726	-1.024
Brandon League	SEA	1.087	1.622	0.534	2.050	0.962

Player	Team	Salary	Full Model	Difference	AIC	Difference
Sean White	SEA	0.415	1.226	0.811	2.088	1.673
Jeremy Affeldt	SFN	4.000	1.354	-2.646	1.725	-2.275
Guillermo Mota	SFN	0.750	1.145	0.395	2.012	1.262
Sergio Romo	SFN	0.416	1.465	1.048	2.839	2.422
Brian Wilson	SFN	6.500	3.872	-2.628	2.266	-4.234
Ryan Franklin	SLN	3.050	2.835	-0.215	3.074	0.024
Kyle McClellan	SLN	0.425	1.160	0.735	2.435	2.010
Trever Miller	SLN	2.000	1.174	-0.826	2.298	0.298
Dennys Reyes	SLN	2.000	1.124	-0.876	1.796	-0.204
Grant Balfour	TBA	2.050	1.143	-0.907	2.577	0.527
Randy Choate	TBA	0.700	1.624	0.924	2.197	1.497
Lance Cormier	TBA	1.200	0.986	-0.214	1.446	0.246
Andy Sonnanstine	TBA	0.417	0.682	0.265	2.490	2.073
Rafael Soriano	TBA	7.250	3.810	-3.440	2.939	-4.311
Dan Wheeler	TBA	3.500	1.619	-1.881	2.347	-1.153
Frank Francisco	TEX	3.265	1.623	-1.642	2.217	-1.048
Dustin Nippert	TEX	0.665	0.847	0.182	1.341	0.676
Darren O'Day	TEX	0.427	1.404	0.977	3.076	2.649
Darren Oliver	TEX	3.000	1.317	-1.683	2.684	-0.316
Chris Ray	TEX	0.975	1.264	0.289	2.145	1.170
Shawn Camp	TOR	1.150	1.307	0.157	2.540	1.390
Scott Downs	TOR	4.000	1.435	-2.565	2.709	-1.291
Jason Frasor	TOR	2.650	1.414	-1.236	1.968	-0.682
Kevin Gregg	TOR	2.000	3.338	1.338	1.630	-0.370
Casey Janssen	TOR	0.700	1.177	0.477	2.439	1.739
Brian Tallet	TOR	2.000	0.785	-1.215	1.540	-0.460
Miguel Batista	WAS	1.000	0.683	-0.317	1.981	0.981
Brian Bruney	WAS	1.500	0.742	-0.758	-0.290	-1.790
Sean Burnett	WAS	0.775	1.551	0.776	2.131	1.356
Matt Capps	WAS	3.500	3.727	0.227	2.470	-1.030
Tyler Walker	WAS	0.650	1.148	0.498	2.698	2.048